Theory and Decision Library A: Rational Choice in Practical Philosophy and Philosophy of Science

Julian Nida-Rümelin

Structural Rationality and Other Essays on Practical Reason

Springer

Theory and Decision Library A:

Rational Choice in Practical Philosophy and Philosophy of Science

Volume 52

Series Editor
Julian Nida-Rümelin
Ludwig-Maximilians-Universität München, Germany

Assistant Editor
Martin Rechenauer
Ludwig-Maximilians-Universität München, Germany

This series deals with practical and social philosophy and also foundational issues in philosophy of science in general that rely on methods broadly based on rational choice. The emphasis in the Series A is on well-argued, thoroughly analytical and philosophical rather than advanced mathematical treatments that use methods from decision theory, game theory and social choice theory. Particular attention is paid to work in practical philosophy broadly conceived, the theory of rationality, issues in collective intentionality, and philosophy of science, especially interdisciplinary approaches to social sciences and economics. Assistant Editor: Martin Rechenauer (München) Editorial Board: Raymond Boudon (Paris), Mario Bunge (Montréal), Franz Dietrich (Paris & East Anglia), Stephan Hartmann (LMU Munich), Martin van Hees (Amsterdam), Isaac Levi (New York), Richard V. Mattessich (Vancouver), Bertrand Munier (Cachan), Olivier Roy (Bayreuth), Amartya K. Sen (Cambridge), Brian Skyrms (Irvine), Wolfgang Spohn (Konstanz), and Katie Steele (London School of Economics).

More information about this series at http://www.springer.com/series/6616

Julian Nida-Rümelin

Structural Rationality and Other Essays on Practical Reason

Springer

Julian Nida-Rümelin
Lehrstuhl für Philosophie und politische Theorie
Ludwig-Maximilians-Universität München
Munich, Germany

Springer International Publishing; Berlin/Dordrecht/Heidelberg/New York, 2019

ISSN 0921-3384 ISSN 2352-2119 (electronic)
Theory and Decision Library A:
ISBN 978-3-319-95506-3 ISBN 978-3-319-95507-0 (eBook)
https://doi.org/10.1007/978-3-319-95507-0

Library of Congress Control Number: 2018956577

© Springer Nature Switzerland AG 2019
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book contains an essay on a conception of practical reason that I call 'structural rationality' and five selected papers published in the course of almost 20 years, discussing different aspects of structural rationality and practical reason in general.

In the meantime, a debate has evolved in the USA concerning structural aspects of rationality,[1] sometimes even using the same term, that I have been using since 1988, when I published my Rejection of Consequentialism both in ethics and in rational choice theory as a habilitation thesis at the University of Munich.[2] I trust that the originality of my conception is sufficiently obvious and therefore I did not add footnotes ex post discussing the differences between my account and these recent proposals.

Every notion of practical rationality has philosophical, even metaphysical, presuppositions and implications. These are touched upon sometimes, especially in part I, but not developed systematically. I hope to do so this in a monograph I am presently working on.

Munich, Germany — Julian Nida-Rümelin

[1] Cf. Suffix of this book.

[2] Cf. JNR Economic rationaly and Practical Reason, Theory and Decision Library Series A, vol 24, Dordrecht/ Boston/ London 1997.

Introduction and Overview

If a theory of practical reason is something like a cookbook containing recipes for how one should act in specific situations, this essay is not a contribution to it. Indeed, the 'cookbook-conception' has led astray philosophy with its insistence on normative criteria. We cannot reinvent reasons for actions; good practical philosophy is not constructivist, neither in the Kantian nor in the Radical sense. A philosophical 'theory' of practical reason cannot substitute our everyday practice of giving and taking reasons; it can only systematize it. One might even say that a theory of practical reasons is impossible: either it is a mere extension of our everyday practical deliberation or it is ipso facto inadequate.

One might call such an approach 'Aristotelian', as Aristoteles argues in extenso in his *Nicomachean Ethics* in favour of phronesis, that is practical wisdom, in opposition to Platonist approaches that take mathematics as the model for practical philosophy. This, however, could lead to the misunderstanding that questions of right acting ('normative' questions) and questions of right judging ('theoretical' questions) have to be addressed by using fundamentally different methods. Both in the history of practical philosophy and in contemporary ethical theory, the model of scientific precision is often contrasted with the model of wise counsel in practical questions. This tendency can lead one to forget that the scientific method is by no means applied to the totality of our descriptive beliefs but only to a small part of them which is relatively unimportant for our lifeworld. Among the descriptive beliefs which are not subjected to scientific analysis – or, at any rate, which are not in need of scientific analysis and justification – are not only everyday observational judgements which are justified by immediate experience but also, for example the realm of the mental, which is constitutive for social life. We take other persons to be agents pursuing intentions, having wishes, sensing pains, etc., and, normally, we can rely pretty well on these judgements. Psychology as a science is quite unimportant for these judgements and their justification. Scientific analysis has a justificatory bearing on our lifeworld only when it comes to conflicting theoretical or practical reasons, that is reasons that speak both for and against a certain descriptive or normative judgement.

Actions can be understood as representing simultaneously two kinds of propositional attitudes, beliefs and desires, based on theoretical and practical reasons accepted by the person. My account differs, however, from the Humean belief-desire model since it takes reasons and not desires as basic. It is reason and not desire that constitutes rationality. This conception of agency does not add reasons to desires or discriminates between basic and derivative desires; it holds the more radical view that agency exclusively represents reasons and that the desires of the acting and of other persons are merely empirical conditions as, for example climate conditions that good practical reasoning takes into account, insofar one might call my account 'stoicist'. According to this neo-stoicist conception, every choice of an action (every preference) is a (normative) judgement (*prohairesis krisis estin*) or at least represents it. This claim goes beyond the traditional polarity of Humean and Kantian philosophy of action, since all actions – not only moral ones – are interpreted as expressions of a normative judgement backed by reasons.

The problem of the transition from a position (as a sort of 'mental attitude') regarding a specific action is not the subject matter of the theory of practical reason but rather should be addressed by a discipline such as neuropsychology. Admittedly, the connection between intentional states and behaviour is relatively obscure; this, however, is not a problem for a theory of practical reason, the scope of which ends in justified intentional states (rational decisions). Basic action-guiding decisions which correspond to basic actions are the fundamental elements underlying criteria of practical reason. The correspondence between manifest behaviour and decisions can be disturbed. Unless such an interference is a case of weakness of will, the agent himself is not accountable. Rather, in such a case, the behaviour no longer qualifies as an action of the person. Mere behaviour, which lacks the constitutive intentional backing and which, consequently, cannot count as an action, cannot be the subject of a theory of practical reason.

The conception dominating today, which I shall call the 'standard view', is often associated with David Hume. Interestingly, this conception misses the specific element of practical reasons, since, in the end, it is only concerned with the rationality of descriptive beliefs. Conative attitudes (wishes, fears, hopes, etc.) can neither be rational nor irrational: they are only subject to rational critique to the extent that they rest on (justified or unjustified) descriptive assumptions. The rational person is motivated by something which is beyond rational critique. In a nutshell, the conception of practical rationality dominating philosophy and several other disciplines today denies that there is genuine practical rationality. In chapter "Acting on Beliefs" of this essay, actions are, in opposition to the standard view, understood as expressing both normative and descriptive beliefs. Normative and descriptive beliefs are equally justifiable. This insight has frequently been associated with unnecessary metaphysical burdens. In fact, there is a far reaching symmetry between empirical and ethical justification of beliefs. In both cases, there is no ontological foundation required. A physical theory is well or poorly justified, and this justification does not depend on the ontological status of the entities with which it is concerned. There is a severe conflict between the standard view and our self-understanding as rational persons. The idea of tracing back an action to motivating (a-rational) desires and

(rational) descriptive beliefs which qualify this action as instrument for the satisfaction of desires is not convincing.

If this diagnosis is correct, then the question arises why the standard view has survived until today. I suppose that it derives its plausibility from a 'physics of wanting' the origins of which can be traced back to the seventeenth century. Hobbesian physics is, of course, very different from present day scientific theories, but its basic assumption that the human body is moved by forces like hate, fear, greed and others, whereas the direction of the movements is influenced by descriptive beliefs, still has its impact on human thought. Rudiments of the scientific developments of earlier times have survived in our everyday knowledge, and they are not without influence on what we deem intuitively plausible. Realizing these influences makes it easier to depart from them.

With that, we will have provided the prerequisites necessary for the conception of structural rationality which will be introduced in chapter "Structures of Agency". Actions are free and rational in virtue of the fact that they conform to certain structures which are intended by the acting person. The person acts 'structurally intentional' according to his form of life without having to abandon his freedom, in contrast to what seems to be required by the standard view. The structurally rational person chooses an act such that it accords to the preferred structure. Additional motives or sanctions creating additional motives are not necessary in order to secure structural rationality.

Chapter "Reasons for Action" is concerned with the structures which impose reasons for action. Reasons for actions are expressed in structures. Being bound to structures is compatible with being free, if these structures are established by reason (a 'stoicist' extension of Kant's categorical imperative). Reasons for action reveal themselves by structures of agency. It is analytically true that we can give reasons for our actions: there is no action without a guiding reason. Inclinations can determine actions only to the extent to which they are accepted by the person as a good reason for action. Inclinations do not directly determine actions but do so via a normative attitude (*krisis*, judgement) of the acting person. Reasons refer to the specific traits of the respective situation, including the subjective traits of the acing person. Reasons, however, are themselves never subjective. Disagreements over reasons for action are to be treated in the same way as disagreements over facts. The form in which we discuss practical questions requires an objectivist interpretation of practical reasons.

A central type of good reasons for action is related to cooperation with other persons. Although there are many different specific kinds of cooperation and although some social practices can only be shown to be cooperative through detailed analysis, there is a simple way of characterizing a general notion of cooperation. It can be shown that individual optimization and cooperation are incompatible when these terms are adequately used. Consequently, genuine cooperation, as is shown in chapter "Cooperation", requires structurally rational actions. Cooperation is a paradigmatic case for the conception of structural rationality: when a person positively evaluates a structure of agency with good reasons, then the individual action which is part of the desired structure eo ipso is rational.

Communication relies on rules. In philosophy of language, the status of these rules is a matter of debate. Intentionalist semantics says that the meaning of linguistic terms can ultimately be reduced to mental states of the communicating persons. One would, however, not be able to understand why people follow the rules constitutive for language, if speaker and hearer were to be regarded as punctual optimizers. In chapter "Communication", it is shown that communicative rationality can only be made fully plausible in the framework of structural rationality.

Our actions are guided by intentions. These intentions display structures, and they are embedded into the overarching contexts of epistemic (beliefs, expectations, assumptions, etc.) and conative (desires, hopes, fears, etc.) attitudes. The rationality of a person is displayed in intentional structures. Theoretical and practical reasons correspond to epistemic and conative attitudes which are based on normative judgements guiding agency (chapter "Structural Intentions").

A person acts out of strength of will when, punctually, she does what she has realized to be structurally rational. Weakness of the will is a case of structurally irrational punctual optimization. Strength of will is a precondition for responsible action (chapter "Weakness of Will").

The conception of structural rationality outlined in this essay can be regarded as an explication of practical coherence. Coherent agency, coherent individual and collective forms of life are structurally rational. In the last chapter, we discuss the role of existential decisions, decisions which themselves can only be rationalized to a limited degree (chapter "Structural Coherence").

Contents

Part I Structural Rationality

Acting on Beliefs 3

Structures of Agency 13

Reasons for Action 21

Cooperation 27

Communication 33

Structural Intentions 41

Weakness of Will 49

Structural Coherence 55

Part II: Other Essays on Practical Reason

Why Consequentialism Fails 65

Why Rational Deontological Action Optimizes Subjective Value 77

Structural Rationality in Game Theory 86

Rationality: Coherence and Structure 98

Structural Rationality and Collective Intentions 112

Suffix

The Meaning(s) of Structural Rationality – A Suffix 127

Part I
Structural Rationality

An earlier version of this essay was published by Reclam, Stuttgart, 2001

Acting on Beliefs

There is a temptingly simple theory of practical rationality which is commonly labelled "Humean" (referring to David Hume) in the literature. I shall leave it open whether this theory really has been put forward by David Hume, which is why I prefer to call it the "standard view". This view is meant to be an analysis of motivation. It is intended to answer the question of what it is that prompts us to perform a certain action. Actions are thought of as the result of some inner motive of the acting person – without this inner motive the person would not act. In order for this inner motive to lead to a specific desire, it needs to be accompanied by certain (descriptive) beliefs. I form the desire to eat a portion of *Spaghetti alla marinara*, since I (1) am hungry (the inner motive) and (2) expect that my hunger will be satisfied by eating a portion of *Spaghetti alla marinara*. Had the belief (2) not been present, my inner motive would have been undirected. This example suggests one further step in the analysis: a distinction between a desire which is solely directed to the satisfaction of hunger and an action-guiding desire mapping out a specific way to attain this goal. The first desire, as opposed to the second, does not direct my agency in any specific way, since a belief as to what would be an appropriate means to satisfy hunger has not yet come into play. With regard to its action-guiding force, the first desire is as yet undirected. In the transition from the feeling of hunger to the adoption of a respective desire, no deliberation is involved as to how this desire might be appropriately satisfied.

In order to make the standard view as plausible and distinctive as possible, we can assume a causal relation between certain sensational states and corresponding desires which are undirected in the sense just introduced. If we assume, like e.g. J. Bentham, that human beings strive at optimizing their sensational states, then the standard view would be bound to a hedonist conception of basic (undirected) desires. This form of hedonism should be understood as strictly empirical (not normative). It does not maintain that a desire is rational insofar as it is directed towards an improvement of subjective sensational states. It rather says that basic desires have been causally brought about in this way. In contrast, as a normative theory,

© Springer Nature Switzerland AG 2019

J. Nida-Rümelin, *Structural Rationality and Other Essays on Practical Reason*,
Theory and Decision Library A: 52,
https://doi.org/10.1007/978-3-319-95507-0_1

hedonism would maintain that one should strive for an improvement of subjective sensational states.

Hungry feelings come and go; for most of us, they are not under (direct) control. Hunger needs to be satisfied in order for the hunger sensation to vanish. For most of us, it is not enough to decide not to be hungry any more in order to make the hungry feeling disappear. For some people it would be desirable not to be hungry as often as they are and, consequently, to lack the desire to eat. Being aware of this fact, however, does not change periodically recurring hunger sensations. This example can be taken to be paradigmatic for the standard view.

Let us suppose there is, stretching out across time, a flow of sensations which are connected to certain subjective emotional states. These emotional states, in turn, are accompanied by a temporal sequence of *undirected* desires which aim at optimizing sensational states. Neither is the flow of sensations over time under our control, nor are we, correspondingly, in control of the respective undirected desires. Rationality only comes into play when descriptive beliefs enter the picture, transforming undirected desires into directed, action-guiding ones. These beliefs can be true or false, and there might be good reasons to believe one proposition to be true and the other to be false. Beliefs alone, however, cannot direct agency, since agency requires a *movens* and this *movens* results from the source of undirected desires which are not in our power and which are, more specifically, beyond the domain of rationality.

The main objection against the standard view is that we are able to give reasons that refer to normative beliefs. Descriptive beliefs may play the role of determining the action best able to satisfy our desires. Whether we then follow the result of this kind of deliberation depends on our normative judgement whether we should choose the respective action, whether it would be good to choose it. Normative beliefs can play an *action-guiding* role, for example in determining the limits imposed on the pursuit of self-interest when considering other persons' interests. Some normative beliefs are not only restricting what we are allowed to do (negative, constraining), but are directly reason-giving (positive, motivating).

Suppose somebody, call her Anna, prefers to see a certain person punished. Anna might back up this preference by saying things such as: "After all, this person is guilty of having acted against the law." Or: "If we did not punish those who have done something of this kind, these crimes would be on the rise." Or: "This is demanded by justice...". Reasons like these can be contested, and full-blown theories are available in the philosophy of law for defending them. There seems to be no way to identify a basic undirected desire that Anna has, that gives reasons to have the preference to see the criminal punished. Anna prefers to see the criminal punished not because she can in this way satisfy some desire she has, but exclusively because she is convinced that this criminal ought to be punished. It is this normative judgement alone that guides Anna's preference. This preference exclusively depends on Anna's normative belief that the criminal ought to be punished and does not at all depend her subjective desires.

Even in cases in which desires are an immediate consequence of emotions, reasons can be given speaking in favour of or against the respective emotions. Suppose somebody says: "I desire that he fails in his pursuit of this project." If he, when

asked for his reasons, replies "Because I hate him", then this emotion in turn is in need of justification. Justifications of emotions generally do not sketch out causal explanations. All of us take certain emotions to be justified or unjustified, i.e., we take it to be appropriate or inappropriate, right or wrong, to have certain emotions when certain circumstances obtain. As our ordinary communicative practices show, emotions are not beyond the scope of normative critique. When pro- and counter-reasons are given, rationality is at stake. The burden of proof is with those who maintain that something which is an object of reasoning in our life-world *nevertheless* is not a proper object of justification, i.e., lacks the possibility of being a matter of rational critique.

The standard view is most convincing when a hedonist anthropology is presupposed. If we as humans were really constituted such that our primary, *undirected* desires were a causal consequence of sensational states and directed towards an improvement of these states, then this would provide a strong case for restricting rationality to theoretical rationality, that is, to the justification of descriptive beliefs. If we distinguish between epistemic and conative states of a person – with the epistemic states being characterized by the descriptive beliefs the person has at that point in time, and the conative states comprising the desires, preferences, hopes (etc.) of this person – a hedonist anthropology offers the possibility to derive all conative states from fundamental, arational, undirected, hedonist primary desires, when the relevant epistemic conditions are taken into account.

The hedonist variant of the standard view, however, can be upheld only when (1) the primary desires, in the way described, result hedonistically from sensational states and (2) all other elements of a conative state, *inter alia* all non-primary desires, can be derived from the primary ones, when the relevant (descriptive) beliefs are taken into account. In order to reject the hedonist variant of the standard view, it would suffice to challenge (1) or (2). As a matter of fact, both (1) and (2) seem completely implausible to me. The causal mechanism by which basic desires are allegedly fixed does not exist, and a reduction to primary desires (independently of how these are constituted) is impossible.

Let us start with (2). What would this reduction look like in the example introduced above? The respective desire would have to result from the fundamental desire under the relevant epistemic premise:

(a) I desire that p.
(b) I know that p is only achievable if q.

(c) I desire that q.

Let us check the plausibility of applying this scheme to the desire that a person is punished for acting in an unjust way. The desire can be expressed thus: (c′) I wish for this person to be punished. What is to be inserted for (a) and (b)?

In (a′), a fundamental desire would have to be mentioned, and (b′) would have to refer to a belief to the effect that punishing this person is necessary in order for the desire to be satisfied. The following instantiations seem straightforward:

(a′) I wish for all those having done something unjust to be punished.
(b′) I am convinced that this person has done something unjust.

From this, it follows

(b″) My desire (a′) that all those having done something unjust be punished is satisfied only when this person is punished.

(c′) I wish for this person to be punished.

The scheme is satisfied. Yet, we cannot be content with this analysis since the structure of motivations is not adequately represented when the origin of the specific desire (c′) is identified with the desire (a′). There does not seem to be any possibility of significantly improving the analysis when staying within the framework provided by the scheme given above. This is so since the desire that the person be punished represents a (normative) belief, namely the belief that people ought to be punished for their unjust doings. It is impossible to reconstruct this motivational structure by invoking a fundamental desire, i.e., by inserting something for *p* in (a). The desire (c′) is justified, or rational, with regard to the normative belief just mentioned, together with the additional (descriptive) supposition that the person in question has acted in an unjust way. The motivational structure can be represented along the following lines:

(a″) Persons (of the type *x*) should be punished for actions (of type *y*).
(b″) This person (is of type *x* and) has performed an unjust act (of type *y*).

(c″) This person should be punished.

Rephrased in terms of the first-person perspective:

(a‴) I believe that (a″).
(b‴) I believe that (b″).

(c‴) I believe that (c″) (and insofar as I desire that the person be punished).

Maybe Anna will only reluctantly agree that she *desires* to see the other person punished. She could insist that she believes the punishment to be appropriate, obligatory or (in a normative sense) necessary, and add that her desires are not at all relevant for this. She could point out the difference in meaning between “approving of the punishment” and “desiring that the punishment takes place” and flesh this out in the following fashion: “There is nothing in it for me if the person is punished – I would not be better off in any way. When my desires are satisfied, however, I am better off than I would be were they not satisfied – simply due to this fact alone, and not because of the realization of something in my interest, which would result from

the satisfaction of a desire of mine. *Ceteris paribus*, the satisfaction of my desires is always in my interest. This is part of the meaning of 'having a desire'. Punishing the person, however, is not in my interest, and consequently there is no respective desire on my part involved."

In order to meet this objection, it is advisable to distinguish between a "desire" (and "desiring") in a narrow, *genuine* sense on the one hand and, on the other hand, a "desire" (and "desiring") in a broader, technical sense. In ordinary language, "desire" is primarily used in the narrow sense. "Desires" in the broader, technical sense are the ones referred to in the brackets of (c‴). Part of the reason for the standard view's suggestiveness is that the two usages of "desire" are not kept separate. The increasing use of the notion of "preferences", it seems to me, is not only due to the influence of decision theory and economics, but is also supported by the need for a single notion which is semantically relatively open and can be used for a wide spectrum of conative attitudes – and by the expectation that the complex plurality of this spectrum can be homogenized by theoretical clarification. I have no problems with the wide usage of the notion of a "desire" (or "preference"). I do not criticize the standard view by denying that, in examples of the type invoked above, a desire that the person is punished results from the belief that the person should be punished. This (wide) usage, however, must not be misused – e.g. as supporting the claim that the desires of the acting person are the basis for all practical reasoning. In our example, this basis is a normative belief and not a desire. This remains the case even though it is trivially true that every action corresponds to desires (in the wide sense).

The person in the example would have been enraged had we imputed that her personal desires were responsible for her belief that the respective person ought to be punished. Surely it is possible for somebody to become angry in the face of an unjust doing, and this anger might cause a desire in the narrower sense, i.e. a desire, the satisfaction of which would positively contribute to his well-being. When the unjust agent is punished he feels better, he is relieved, his rage blows over. In this case, the transition in (c‴) from "I believe that this person should be punished" to "I wish for this person to be punished" is of a different sort. Doubtless, in this case it is a genuine desire and not only a desire in the wider, technical (and ultimately trivial) sense. This, however, does not affect the fact that this desire is also grounded in a normative judgement, and, hopefully, the one who is enraged makes his desire for the other person to be punished dependent on the fact that the punishment is justified.

Obstinate adherents to the standard view could relapse at this point by insisting that the transition from the singular normative belief to the respective desire in (c‴) would be a *non sequitur*, which is why even (a‴) would have to be phrased in terms of a desire rather than a belief. Surely, there might well be people who ground a specific desire for a certain person to be punished in a more general desire. This, however, is far from necessary, and in the example under discussion it would be rather unusual for somebody who is not a sadist. A rational person's desires correspond to her normative and descriptive beliefs. A normative belief does not lead to the corresponding desire in a way which can be called "logical" or "causal". Our

claim is weaker: it says that a rational person with the respective normative and descriptive beliefs has the corresponding desire (at least under ordinary circumstances). This claim is part of the characterization of a rational person.

Examples of the kind discussed above show that the standard view clashes with our understanding of ourselves as rational persons. It might be the case that this self-understanding rests on an empirically false theory which we have adopted and which guides our everyday activities. In case of such a striking conflict with some of our most basic life-world beliefs, however, the burden of proof is with those who believe in the *standard view*. Early in the twentieth century there might have been hope for psychoanalysis to be able to carry this burden of proof. Today, at any rate, this hope seems rather naïve.

Even if there were basic, undirected desires and even if these desires really were of a hedonist kind, the situation would not be any better. Rather, the crucial deficiency of the standard view can be made especially clear under these circumstances which seem so favourable for this theory. If the basic desires are hedonistic, then the degree of desire-satisfaction over time suggests itself as the measure of practical rationality. If the basic desires are hedonistic, it is plausible to consider someone's action to be rational to the extent to which it is in accordance with his basic desires. Under this assumption, there is a fundamental objection against the *standard view* which I shall call the "argument from intertemporal coordination" and which I explain in what follows.

From our everyday practice we know that prospective agency is often better than the pursuit of momentary desires. After breakfast, for instance, I am not hungry. I know, however, that I will be hungry again in a few hours which is why it makes sense to do some shopping in order for me to be able to satisfy my hunger in several hours from now in an appropriate way. Presently, I am not hungry, and I do not experience any need for going shopping. I go shopping solely due to my considering the fact that I expect the latter situation to occur in which I will have a desire which I am now lacking—a desire I will only be able to satisfy if I take appropriate action (i.e., shopping) now. The adherent to the standard view can insist that, obviously, I have a desire for shopping now even though I am not hungry, since I have, among others, the desire for being able to satisfy the hunger I will experience in the future. And I would not deny this. Of course it is possible to say that I have the *desire* for not being (or, rather, not remaining) hungry later on. This desire, however, results from the insight that already now – and independently of my present desires – I have a *good reason* to do some shopping in order to be able to cope with the desire I now have reason to expect to have in the future.

Prospective agency is an obvious and, as it were, natural component of practical reason as it occurs in everyday life. A theory of rationality which does not leave room for genuine prospective agency seems to be bound to fail right from the start. Although I am in good company here, a certain reluctance to accept the point just made is understandable. How can it be that the outlook on rationality, which I have called the "standard view", is unable to accommodate a phenomenon which is so crucial to our understanding of rational agency? As a matter of fact, the standard view can be immunized against the objection by interpreting the prospective stance

towards future desires which one has reason to expect to have as a present desire in the narrow sense. A strict refutation of the standard view, therefore, is not possible along these lines. At any rate, however, the artificiality of the manoeuvre just presented is striking enough. Hence the survival of the standard view seems to depend on whether there is an alternative, theoretically more convincing integration of this element of our life-world practical rationality. The conception of structural rationality for which I argue in this work is meant to be such an alternative.

It is useful to frame the argument from intertemporal coordination in a somewhat more abstract way. Let us imagine an agent has a number of options to go for at either of two points of time, T_1 and T_2. Now suppose the following holds: when the person acts according to her respective basic desires at each point of time, the sequence of actions $<A_1, A_2>$ would result. (We shall call an N-tupel of acts which are ordered according to their temporal sequence a *temporal action structure*. If there is only one person involved, we are facing an *intrapersonal temporal action structure*.) It is compatible with this assumption that another intrapersonal temporal action structure $<A_1', A_2'>$ would satisfy the same basic desires to a greater degree. In general, temporal action structures resulting from punctual optimization are not optimal. (Punctual optimization is to be understood here such that, at each point in time, the action optimizing the present desires in the narrow sense is chosen.) The rational person will choose her actions at any point in time such that they can be embedded into a temporal action sequence which she generally approves of – i.e., taking into account her expectations about future decision situations. The difference between punctual and structural rationality just described is the starting point for the exposition of the view to be defended in this book.

For a theory of practical rationality to be adequate, it needs to be able to allow for the fact that specific intentions are rendered rational only with regard to other, more general intentions. While the standard view allows for a high degree of contingency by making rationality dependent on momentary desires which cannot be further justified, the Kantian theory attempts to determine rationality as independent of contingencies, i.e. of the agent's desires and 'inclinations'. The conception of structural rationality is situated between these two extremes. Kantians will believe it to be too 'Humean' in spirit, since 'inclinations' play a constitutive role for rationality, while Humeans might find it too 'Kantian', since its heart is the idea of conformity with rules. The central role which structural aspects play for every rational practice renders plausible a lasting coexistence of Humean and Kantian theories (insofar as both capture a crucial element of practical rationality, which needs to be taken care of in an adequate theory).

From the perspective of the person having the desires it is irrational at each point in time to satisfy the respective momentary desires. Prudence – understood as intertemporal and intrapersonal coherence – requires a certain distance from the desires that are present at a specific point in time. Punctual optimization of my desires results in a degree of desire-satisfaction which is lower than the one which is possible if I were to dispense of punctual optimization. In virtue of my knowledge of this, I refrain from the punctual satisfaction of my desires.

The most natural interpretation of this manoeuvre, it seems to me, is not to assume a shift in the structure of my desires. Rather, the mere (descriptive) knowledge of the fact just described causes a change in my motivational situation. Due to the structural irrationality of punctual optimization, I dissociate myself from the given structure of my desires. This knowledge itself determines the motive for action. It seems rather implausible to make the standard view compatible with such a change in the motivational situation by assuming a respective change in the structure of desires. My desires (in the narrow sense) have not changed at all. Rather, the object of my evaluation has been broadened – from punctual accessible individual actions to intertemporal *action structures*.

If one wishes to defend the standard view against this argument, one might insist that the change in the motivational situation has obviously resulted in a change of certain desires at the given point of time, since a desire is ultimately realized by the respective action. Even though this manoeuvre seems to be elegant at first sight, it is, when carried further, devastating for the standard view. This is so since every contrast with other theories of action (e.g. the Kantian one) vanishes if those desires which are the result of rational deliberation are incorporated into the basic model of the standard view. Ultimately, the Kantian respect for the *Moral Law* also results in a desire to act on maxims that pass the test of the categorical imperative. The standard view differs from the Kantian approach in assuming that motives for action have to be grounded in something which lies beyond rational control. Desires which are admissible within the standard view can be determined in a recursive way: the reduction of a desire D' to a desire D amounts to determining those descriptive beliefs E for which it holds that $D \,\&\, E \rightarrow D'$. The standard view must postulate that for all desires there is a chain of reductions the end of which is a basic desire that is 'given' and not itself the result of deliberation. A desire that is based on normative beliefs like Anna's conviction that "this criminal ought to be punished" or Kant's respect for the Moral Law cannot be recursively reduced in this sense.

If people do in fact have desires which cannot be reduced to basic desires (i.e. desires which are beyond the scope of rationality), then the standard view, understood as an *empirical* theory about motivations, is false. If our understanding of what it means to be a rational person cannot do without the assumption of desires which cannot be reduced to basic desires, then the standard view, understood as a normative theory of good practical reasons, is false. It seems to me that both of these statements are true.

It seems to me that the standard view owes lot to an implicit *physics of willing and acting*. In the writings of Thomas Hobbes, a 'physicalistic' picture of agency is explicitly worked out. Love and hate are understood in analogy to attractive and repellent forces, and they keep the human body in motion. Agency is a form of movement. This is why agency requires the operation of some force, which is brought about by affection and aversion and, in general, by passions or inclinations. The cognitive element, our perception of and our judgements about the world, remains motivationally inert. Considered alone, it cannot bring about movements and, consequently, it cannot cause actions. This background theory lives on in our common-sense conceptions of the world, as do other rudiments of the scientific

development of earlier days. As the history of science shows, it often takes centuries until new scientific results are finally accepted as 'common sense,' and conversely, the old and scientifically outdated theories often seem to have a high degree of common-sense plausibility. When such a background theory is being made explicit, one can generally assume that it has already been refuted. Matters are different, however, with regard to successor theories, which are inspired by those theories and which hide their theoretical roots in a more or less sophisticated fashion. In this chapter I attempt to unpack the line of thought responsible for the survival of the *physics of willing and acting* that actually stands in the way of an adequate understanding of human agency. In so doing, I shall make reference to another element of the conception of structural rationality: *agency without impetus*.

Let us call the 'physicalist' picture of human agency the *impetus theory*. Agency is movement; it is an intervention in the causal story of the world, and it can, consequently, only be made sense of if we assume appropriate moving forces that are to be located in the conative states of the agent. No impetus, no action. An action surmounts, as it were, inner and outer resistances, and in order for the action to be able to do this, some impetus is required. This impetus is given a direction by descriptive beliefs. It was already Hobbes who took the faculty of foresight (*prospectus*) to be characteristically human. Foresight presupposes knowledge about empirical regularities which, however we acquired it, helps us in orienting our agency in the world. In close analogy to mechanics, the gradient of energy potentials corresponds to the respective gradient of hedonic states (the level of subjective well-being). The way in which the latter gradient leads to movement (i.e., to actions) depends on the empirical and epistemic circumstances of the situation. The belief that a certain course of action will not contribute to the satisfaction of desire is comparable to a wall obstructing movement. Under favourable circumstances, i.e. in rational and suitably well-informed agents, empirical restrictions will be anticipated in action-guiding beliefs.

This physics of willing and acting would make sense if we had no physical theory of which we have reason to assume that it is, at least in principle, able to cover all physical aspects of human agency. Surely, a complete description would have to be overwhelmingly complex, the best computers in the world would not be able to handle it, and it is doubtful whether our knowledge would suffice to describe the relevant processes completely even if we had sufficiently strong calculating capacities, since many neurophysiological processes are not yet well-understood. At any rate, we know how such a description would look like in principle, on the basis of the physical (and biochemical) theories available to us today. Such a complete description would not have to comprise any reference to conative states (such as, e.g., desires). In order to arrive at such a complete description, there is no need for a duplication of the physical description on the level of mental states. Fields of forces, energy potentials, molecular structures, neurophysiological processes, muscular contractions and external conditions etc. are, at least in principle, sufficient for the explanation of bodily movements, all within the conceptual frame of the natural sciences. There is no gap which would have to be filled in by the assumption of an impetus provided by conative attitudes. And if such a gap were to occur (a problem

which we shall come back to in the next chapter), there is no reason to assume that it can be closed by a quasi-physicalist theory.

There is ample reason to tie actions conceptually to movements of the body. Contemporary physics is available as a scientific theory for the explanation of these movements. Philosophers should not strive to replace or amend it by some other, rather obscure theory going back to a time at which modern physics had not yet been developed. Hence, the analytical tools provided by the modern sciences suffice in principle to explain the movements and the bodily processes associated with human agency. The question remains whether a psychic analogue to physical forces and energy is required on the level of the analysis of mental and, especially, intentional processes. The impetus theory introduced above provides such an analogue. Our rejection of the impetus theory rests on the assumption that there is a far-reaching analogy between theoretical and practical reasons.

Let us look at the way in which so-called 'theoretical' reasons operate with regard to a mathematical example. Suppose somebody puts forward a mathematical assertion *M*. He is asked: "Why do you believe *M* to hold?" The one who was questioned could sit down, grab a piece of paper and (without speaking) sketch a proof for *M*. If the interrogator is convinced that the sketch is correct, he will not inquire further as to why the colleague is convinced that *M* holds. In this way, *M* is justified. Reasons for *M*, or, to be more precise, for the truth of *M*, have been given. In general it would be incomprehensible for us if the interrogator – convinced that the sketched proof of *M* is valid – nevertheless continued to ask why his colleague believes in *M*. If somebody possesses a valid proof of a mathematical statement *M*, then he has good reason to believe in *M*. There is no need for a further reason –e.g., that it is in an individual's interests to believe in M, or a sense of duty to integrate mathematical assertions into one's set of beliefs – provided such a proof is available. We regard 'theoretical' reasons as *self-sufficient* in this sense. They suffice to motivate an appropriate modification (extension or revision) of our system of beliefs. What is more: we believe a person to be irrational when she has theoretical reasons for a modifying her system of beliefs without being motivated to a respective modification. There is no need for an additional motive in order for theoretical reasons for our beliefs to become effective.

There is a difference between changes of beliefs and actions insofar as actions are tied to externally observable behaviour. Only radical behaviourists believe the same to be true of beliefs. Changes of beliefs occur – as one might think – without effort, whereas actions do require effort and a respective motive as impetus. But in fact there is nothing to be overcome by psychic energy, since bodily movements associated with actions belong to the sphere of natural facts, the description of which is a matter of the natural sciences which lack the presupposition of mental and intentional entities. If there is a need for an impetus, this impetus is provided by physics. Intentions which are constitutive for actions do not require an impetus but rather (practical) reasons. The impetus theory of action and the idea of inert epistemic states, on the one hand, and the idea that desires drive actions on the other, should be dismissed completely.

Structures of Agency

A rational person is author of her life: she is imposing structures on her agency. She pursues long-term projects that are realizable only if a sufficiently large part of her actions is aligned with these projects. Her moral convictions, the moral rules she follows, the moral values she has, her virtues reveal themselves in structural traits of her agency. It will hardly be disputed that structural traits of the kinds mentioned do in fact play a major role in our agency. Yet, how can these structural traits be accounted for in a conception of practical rationality? There are two answers to this questions which dominate the philosophical scene.

The first answer says that structural traits are external constraints on rational agency. These traits are thought of as constraining the range of options available to the acting person. It is compatible with this interpretation that certain structural traits can be freely chosen, as, e.g., in acts of self-binding in which a person deliberately decides to restrict the set of alternative actions. Many structural traits do indeed play an 'external' role, so understood. Among them are not only the genetic characteristics which shape our agency, but also some of the effects of education and socialization. The role of structural characteristics is external when individual actions conforming to these characteristics are no longer free with regard to this very conformity – i.e., when the person could not choose a non-conforming action. Even in those cases in which the structural traits have become well-established over a long period of habituation, such that each instance of conformity with these structural traits is no longer a matter of decision, it would be wrong to interpret this phenomenon as a result of a determination external to the acting person.

Psychologists often insist that most of human practice is unconscious. Sometimes it seems that there is a deep (paradigmatic) conflict between psychology and (practical) philosophy, psychologists being reason sceptics, philosopher being reason adherents. It seems to me however that this apparent conflict rests on conceptual confusions that can be overcome if one uses the conceptual frame of structural rationality. Take for example a person who learns how to drive a car. If she decides to turn left at the second crossing to come, she has a lot to decide: when to slow down,

© Springer Nature Switzerland AG 2019

J. Nida-Rümelin, *Structural Rationality and Other Essays on Practical Reason*,
Theory and Decision Library A: 52,
https://doi.org/10.1007/978-3-319-95507-0_2

when to look in rear mirror, when to turn the blinker on, when to change gears etc. The experienced driver in contrast, does all this automatically, she just decides to turn left at the second crossing to come, and then she turns left, All what she has to do in order to get that result is unconscious or automatic. The experience in driving shows itself in less decisions to be taken, the control the expert driver has over his behaviour is even augmented when the number of decisions to be taken in order to get the same result diminishes. In terms of structural rationality: authorship of agency reveals itself in controlling the structural traits of ones behaviour, the 'bigger' these structures become the more souvereinty of agency there is. The account of structural rationality offers a moleculare (holistic), instead of an atomist (particularist) interpretation of human agency, as the standard account does.

In an adequate theory of practical rationality, structural traits have to be interpreted such that they are compatible with the individual's agency and authorship. If this does not hold for some structural traits, if these structural traits do indeed causally constrain the range of options available to the acting person, then they can no longer be considered to be part of the action. Only those parts of a person's behaviour which are under the acting person's control can be understood as genuine actions. This is not the case with external causal constraints. Hence, insofar as structural traits of agency represent external constraints, they alter the causal circumstances of agency, but are not part of the action itself.

The second answer to the question – how can these structural traits be accounted for in an adequate conception of practical rationality? – is Kantian: Freedom of action manifests itself precisely in those laws of action which the acting person imposes upon herself. According to Kant, free actions are conceived of as being guided by moral reasons, i.e., by respect for the Moral Law. Actions which are not guided by moral reasons are not considered free, since they are determined by pragmatic and technical imperatives; in other words, they are an immediate consequence of contingent desires. At the transition from technical to pragmatic imperatives, however, it should already become clear that, even within a Kantian framework, only a dissociation from desires operating at their respective points in time allows for the realization of pragmatic imperatives, i.e., individual happiness. Without genuine freedom of action—without the ability to act according to self-imposed rules—the pursuit of happiness is futile. The problem of *binding* actions, therefore, does not only arise in the context of moral agency, but generally with regard to intentional actions aimed at the realization of goals, the scope of which lies beyond the time of action. Only if humans were 'punctual' beings, meaning that our epistemic and conative states would not display any intertemporal structure, would acting purely out of desire (and the epistemic state at the given time) be conceivable at all.

The difference between the Humean and the Kantian approach only holds with regard to the moral case, while both theories interpret non-morally motivated actions as the satisfaction of punctual desires (i.e., those desires bound to the respective points in time at which they occur). Yet the Kantian critique of the Hume's approach with regard to moral actions is relevant for actions of any sort. Conditions of agency arise, among other things, from my being aware of the fact that the satisfaction of my present desires would rule out the satisfaction of desires I expect myself to have

at some later point in time. Conditions of agency stem from our duties to others, and they are grounded in rules which I have adopted since their acceptance does not restrict the freedom of my punctual agency (the respective momentary decision).

There are many different forms of action binding. Some of them can be reconstructed by optimizing the consequences of my actions over time in a general way, an approach which normally requires me to refrain from punctual optimization. Other ways of action binding are not consequentialist in character. This is especially true when the relevant reasons for action are non-consequential:

1. Commissiva: I have a reason to keep my promises, independently from the consequences of keeping the promise
2. Libertates: I have a reason to refrain from doing something what would interfere with rights and liberties of some person
3. Officia: I have reason to help my closest family members if in need, even if there are better possibilities to optimize universal well-being

to mention three important types of practical reasons, that gave rise to opposed ethical theories: contractualism, libertarianism and communitarianis. Each of these theories taking one of these types as basic and dismissing – inadequately it seems to me – the other types as normatively irrelevant. Most of our practical reasons do not refer to the optimization of consequences. Having promised something is a good reason for acting in a way that fulfils the promise. This reason for action is independent of whether fulfilling the promise has optimal consequences. The promise itself constitutes a good reason to do something. Of course, this does not mean that considerations about the consequences play no role whatsoever when I deliberate whether I ought to keep the promise. Surely the keeping of a promise can cause disadvantages so severe that, all in all, it seems obligatory to break the promise. Here it is important to realize that the fact that I have made a promise constitutes a reason to act in a certain way, even if this reason may in some cases compete with other reasons, also relevant in the given situation.

In contemporary practical philosophy on both sides of the Atlantic Ocean, the view is prominent that there are two basic forms of rationality – 'strategic' and 'communicative' rationality, to put it in the terminology of discourse ethics. Strategic rationality corresponds to consequential reasons and communicative rationality corresponds to non-consequential reasons for action. The point of this distinction, however, lies in the fact that – in the tradition of Kantian thought – it establishes a dichotomy which cannot be maintained if our argument is correct. It is also true of 'strategic' rationality that it is successful only if it embeds the individual's punctual action into a structural context. Communicative agency is a special case of cooperative agency, and without cooperation, strategic agency would not be successful in the long run. In a separate chapter "Communication", I intend to show that for any form of cooperation – not only cooperation constituted by communicative agency – it is necessary that (cooperative) rules be sustained. For individual agency to be in conformity with these rules, it must go beyond punctual optimization. The atomist and consequentialist agent is unable to cooperate and, more specifically, unable to communicate.

Most forms of interaction depend on the fact that the agents act in accordance with certain rules – consciously or unconsciously. These rules range from conventions of politeness or reliability to overarching rules of considerateness for the interests and the evaluative attitudes of others. The respective forms of interaction depend on a sufficiently high degree of conformity with these rules. This does not exclude the possibility that agents act against these rules and thereby exploit an established scheme of (cooperative) interaction, a form of parasitic behaviour. This type of behaviour, however, needs to be distinguished from types of behaviour, where the decision to act against rules that are constitutive for a certain form of social interaction is motivated by normative critique. Such a critique, however, is only possible on a 'local scale', a global critique would put the persons outside of the structure of social interaction, they would loose their status as interactors, as socially interacting members of a society or a community. Not all rules of interaction are normatively binding structures of agency. The basis of a normative critique of etsablished structures of agency, however, can never come from a point of view which is external in that it does not refer to any accepted rules of interaction and cooperation. There is no Archimedian point of normative critique outside.

Nevertheless, even if one focuses one an individual's pursuit of her project, rational agency devoid of structures is not conceivable. A person without recognizable structures of agency would appear unsteady, desultory and incoherent, and one could not assign to her aims and attitudes constant over time. Reasons for action would, for such a person, lose their relevance from one point of time to the next; others would be unable to interpret her actions as revealing certain values sufficiently stable over time, and it would be impossible to assign a utility function – in the decision-theoretic sense – to her preferences.

Since it is true of every person that wishes and inclinations change, there are problems of coordination with regard to the actions of an individual person: difficulties which are analogous to interpersonal problems of coordination. What determines the interaction of several persons at a given point in time can be mapped onto the individual with regard to the coordination of her actions at different points in time. When we, additionally, take into account the fact that a rational person can be confronted with conflicts occurring between different conative attitudes at a single point in time, then the possibilities of interaction which may occur in an *interpersonal* setting can be transferred to the *intrapersonal* case.

To put the matter in economic terminology: transferring interpersonal structures of interaction to practical conflicts of decision within an individual person amounts to recognizing an intrapersonal analogue to the problem of collective goods. In the interpersonal case of collective goods, individual optimization leads to a suboptimal situation, i.e., collective goods are not provided for to a degree which satisfies the needs of those involved in the interaction. The allocation of resources is inefficient. In societies, the problem of collective goods is partly solved by state intervention, e.g., by financing institutions of education and homeland security using tax money, and partly through civil engagement. While the first answer to the problem of collective goods is largely compatible with individual optimization on the part of the citizens, this is not the case with civil engagement. Actions of this sort should be

interpreted as a form of voluntary bonding to common projects, which imposes constraints on individual optimization. Motivated citizens spend part of their leisure or income providing collective goods or services, although the individual's voluntary action does not typically contribute to the provision of collective goods in a way that is recognizable to the individual himself.

Due to the structural analogy between situations of interaction in the inter- and the intra-personal case, the problem of collective goods occurs in the latter as well. The person has to refrain from punctual optimization in order to best satisfy her overall needs (a multiplicity of value functions representing the internal practical conflicts of the person). This economic description of the problem is not wholly unproblematic: the economic analysis of collective goods treats conformity to rules in an instrumental way and therefore remains within the scope of consequentialist thought, broadly conceived. At any rate, it makes visible the discrepancy between individual optimization and collective rationality.

Structuring actions over time is hence a necessary precondition of rational agency both in the inter- and in the intra-personal case. When a group of persons facing a collective goods problem has the opportunity to do so, it imposes upon itself certain restrictions – for example, via the introduction of sanctions – which ensures that the collective good is provided. The smaller the group and the higher the degree of moral motivation, the less need there is for external sanctioning. There are, however, many social phenomena which show the existence of moral motivation even in large, anonymous groups. One of these phenomena is that the majority of democratically-minded citizens take the effort to vote, although this act is neither rewarded, nor are non-voters punished, and although all are aware of the fact that the probability that the individual vote has any influence on what will actually happen is practically zero. In this case voting is irrational for economic orthodoxy while structurally rational.

If it is true for a group of N persons that each of N persons of a group prefers a situation in which all group members follow some rule R which enables the provision of a certain collective good under conditions of mutual trust it is structurally rational to conform to this rule even though no sanctions are involved. Any theory of practical rationality which declares a cooperative attitude of this sort and the corresponding readiness to conform to the rule as irrational cannot be adequate. A theory of practical rationality must allow that rational persons voluntarily conform to a rule (provided they are motivated appropriately), the sanctioning of which they would rationally agree to. Why shouldn't people do something without enforcement which they would do when forced to do so, if they agreed to the enforcement? Again: this argument does not preclude that it is rational for groups facing collective goods problems to impose external sanctions. There is, however, no reason why adherence to rules which are in everybody's interest should only be possible under external coercion and not also be motivated by an insight into the structure of the situation of interaction and the common interest in cooperation.

The same is true with regard to the intrapersonal case. Insight into the expediency of conformity to a rule that would solve the intrapersonal collective goods problem can be sufficient to motivate the structurally rational person to act

accordingly. It is not necessary to assume additional motives which prompt the person to choose the rule-conforming action. The action is chosen by virtue of its being part of the desired structure. The action is not chosen because it corresponds to a present desire of the person. It is this feature which one might find most closely related to the Kantian conception: the reasonable choice of actions is autonomous and not heteronomous, i.e., determined by momentary inclinations. This similarity, however, should not lead to normative apriorism: the desire to realize a certain structure of agency is not at all independent from inclinations, but is reasonably formed with regard to the background of the whole complex of conative attitudes of the person. Moral categories in the traditional sense (consideration for others, etc.) as well as other inclinations (basic desires) must be included. Attempts to formalize this process are bound to fail. There are no logical or quasi-logical criteria on the basis of which the rationality of structures could be assessed. In this respect, the practical philosophy of Kant and of many of his successors is connected to an untenable form of rationalism. We remain trapped in the complex of our conative and epistemic attitudes, and an improvement with regard to practical rationality comes in the moderate form of a greater degree of coherence of these two interconnected types of attitudes.

We call an action "structurally intentional" if it has been chosen with regard to its conformity to a structure: the person intends to realize a certain temporal structure of agency and she chooses this action for that reason (= accordingly motivated). Often, it will be difficult for the person to make explicit the intended structure of agency. This, however, does not affect the claim that this form of motivation for action is not a special case, but rather a normal one. In striking contrast to the assumptions of the standard view, an action that comes about without any structural intention is hardly conceivable. The examples now and then discussed in the literature on the standard view typically lack appropriate embedding into a justificatory context and are inasmuch only incompletely described. As soon as the action-guiding reasons are made explicit, the appearance of punctual intentionality cannot be upheld. In the next chapter, we will discuss the role of reasons for action within the conception of structural rationality. For purposes of theoretical analysis, we may, of course, legitimately isolate simple examples of actions from their wider context of guiding reasons, conative attitudes and epistemic beliefs. An isolation of this kind, however, must not lead to the tacit establishment of a paradigm which is sometimes called "intentionalistic" and which is bound to be blind with regard to structures.

Here is a well-known example:

(1) A person *P* wishes to heat the hut;
(2) *P* knows that in order to heat the hut, it is necessary to light a fire.

From the conative premise (1) and the epistemic premise (2), the rational action is inferred:

(3) *P* lights a fire.

Understood as a paradigm, this example can lead to the misunderstanding that practical rationality – as discussed in chapter "Acting on Beliefs" – consists in conforming to momentary inclinations (basic desires) through an action which is directed by certain (epistemic) beliefs.

The rationality of an action cannot be determined in a punctual fashion. This is even true of simple isolated examples of this kind. Rationality always, either implicitly or explicitly, refers to structures. The person who is lighting the fire, for example, has to make sure that she does not hurt anyone in doing so, that lighting fires in situations of the kind she is just facing would not lead to ecological catastrophes when acting in this way is adopted as a general rule. In other words, she has to make sure that acting on her momentary inclination to heat the hut really constitutes a good reason. This can only be done by embedding her punctual attitude towards the decision intoto a wider structural setting. Generally, this embedding works by way of restricting the available options. This is why structures can be largely characterized using the language of controlling and restricting agency. Structures are realized by way of a restriction of the respective punctual optimization. This is why, *a limine*, one might talk of pure "structural intentionality" as well as of pure "punctual intentionality." Purely structurally intentional actions are solely determined by their respective structural reasons: reasons which refer to the fact that the action conforms to the respective structure. When more than one option to which this applies is available, there is still room for the concept of a (purely) structurally intentional action provided that the person, with regard to her momentary conative and epistemic attitudes, is indifferent to these options. Hence, the presence of a (purely) structurally intentional action does not require that there be only one action conforming to the respective structure. The (purely) punctual intentional actions would be those in which binding and restriction by intended structures does not play any role at all. These two pure forms are ideal types of agency that do not occur frequently in everyday life. A widespread hybrid form of agency optimizes punctually within the boundaries set by intended structures.

Not every (temporal) structure of agency is the result of structurally intentional actions. In many cases, temporal structures of agency result from external constraints which leave the person with no other option apart from acting in conformity with the respective structure. Consequently, we need to distinguish between those (temporal) structures of agency which are realized by actions conforming to the respective structures and which are freely chosen by the person, and structures for which this is not the case. We call a structure of agency "intentional" if the individual acts out of a decision to conform to the structure. A structure of agency is intentional if its constituent actions are structurally intentional. Structures arising from punctual intentional decisions are not intentional, since their occurrence is, as it were, only a side-effect of punctual decisions. Finally, one can distinguish between "explicitly intentional" and "implicitly intentional" structures of agency. A structure of agency would be implicitly intentional if the person were to be unaware of the rule on which she bases her decision to act, while it remains true that her individual actions are freely chosen, for example, with regard to a specific conative attitude, which leads to the respective rule conformity. Structures of agency are explicitly

intentional, if the person consciously chooses her actions with regard to the respective rule. This corresponds to the Kantian model of rational motivation. If a structure of agency does not result from structurally intentional individual decisions, we can express this by saying that the structure is "non-intentional".

Here is an example to illustrate these distinctions. Suppose that a smoker comes to realize that it would be better for him to stop smoking. In order to do so, he would have to abandon the habit of lighting a cigarette in certain situations. With regard to his future actions, he subscribes to the simple rule not to light a cigarette, come what may. His spontaneous desires, however, which are not under the control of rational deliberation, point in another direction; they would be satisfied by lighting a cigarette in those typical situations in which these desires normally occur. Fearing that the strength of his will might not suffice, the smoker might decide to live in a solitary mountain refuge for a couple of weeks where no cigarettes are available. He can realize this plan and thereby impose an external structure of agency upon himself which, in time, stops being intentional. Once the decision to go to the mountains is made, he can no longer deviate from the rule not to light a cigarette. If the smoker, however, trusting the strength of his will, stays at home and nevertheless manages to resist a multitude of temptations to light a cigarette, then this structure of agency is (explicitly) intentional. The respective acts and omissions (not lighting a cigarette in the respective situations despite the presence of a spontaneous desire) are in accordance with the rule, and this corresponds to a standing intention of the smoker (or rather, the ex-smoker): the individual decision is structurally intentional. Surely, with regard to the smoker who goes to the mountains, one can also say that the intention to stop smoking persists. The ongoing adherence to the rule not to light a cigarette, however, is not the result of free (structurally intentional) individual decisions.

The person believing that she will manage without external constraints trusts her strength of will, which enables her to act against momentary inclinations dominating at specific points in time. In every individual case in which she refrains from lighting a cigarette despite the presence of strong inclinations to do otherwise, she does so in virtue of her intention to stick to the self-imposed rule and, thereby, to actualize the respective structure of agency. It would be artificial to interpret this long sequence of conscious omissions (not lighting a cigarette) as the result of a process of deliberation in the individual case. It is not true that the unwholesome consequences of a further cigarette stop one from lighting another one. Rather, it is the intention to stop smoking once and for all (i.e., to adhere to the rule which adherence is constituted by a sequence of omissions). Even if, under a punctual description of the situation, the momentary deviation from the intended structure of agency is taken into account, it would not correspond to the structure of her motivations in such a situation to regard the individual omission as a result of weighing consequences and side-effects. The structure of motivation is, in the first place, characterized by the intention to adhere to the respective rule and not to follow momentarily opposed inclinations. Even when enriched, e.g., by intrinsically valuing rule-conformity and similar tricks, the punctual description of the case remains unconvincing. It is only the "structural" view which leads to a correct description of the situation by introducing structurally intentional decisions into the analysis.

Reasons for Action

Structures are imposed on agency by external as well as internal factors. Some of the internal factors are of a rational sort: the person accepts a certain reason, and this is revealed by his structuring his actions accordingly. Reasons are external – according to the position defended in chapter "Acting on Beliefs" – they are independent of the subjective state the person might be in, they are also independent of collective subjective states of groups of persons, culturally, ethnically, nationally, religiously or otherwise defined. Reasons, however, become action-guiding only when they are accepted as reasons. Accepting a reason belongs to the internal, subjective factors, although reasons themselves are not part of the person's subjective states.

Consider an example. Many things we do in order to be considerate to others. "Why did you do that?"— "I did it because I wanted to be considerate." Such a response might be in need of further explanation; yet, it cannot be doubted that the duty to be considerate can count as a good reason for doing something. Considerateness and the basic attitude of respect for other persons and for the forms of life chosen by them (their specific desires and needs, their evaluative attitudes) result in certain structural constraints on options for acting. Partly, these structural constraints are laid down in positive laws and are sanctioned accordingly. To a much greater extent, however, they have to be determined again and again by each person, and it requires a certain amount of empathy and sensitivity to determine the corresponding rules. Here, the virtue ethics approach tends to be sceptical about explicit rules and rather relies on the appropriate attitude (virtue) of considerateness and respect. Under favourable circumstances, i.e. when the strength of will is sufficiently high and the assessment of the situation is cognitively appropriate, the accepted reasons for action are revealed in our actual choices. The accepted reasons for actions impose structures on our agency. To accept a specific reason for action (and to make this reason explicit when faced with critical questions), ties the agent to the respective structural trait of his agency. A person who expresses a reason for action without displaying the corresponding structure in her agency loses credibility

© Springer Nature Switzerland AG 2019

J. Nida-Rümelin, *Structural Rationality and Other Essays on Practical Reason*,
Theory and Decision Library A: 52,
https://doi.org/10.1007/978-3-319-95507-0_3

in the extreme case she becomes as a person incomprehensible: you do not know any more who she is.

Proponents of the standard view could object that actions express inclinations rather than the acceptance of certain reasons for action. With regard to this manoeuvre, I have already said a great deal in the first chapter. Here, we only need to add that our ability to give reasons for every action we perform speaks against the assumption that inclinations directly determine actions. It is analytically true that we can give reasons for actions. If we cannot give reasons for some of our behaviour (but, at most, causes), this behaviour is not to be characterized as acting. Actions are consciously chosen, and accordingly persons who are sufficiently mentally equipped can give those reasons which guided their actions (or decisions). In the cases of those individuals who are less intellectually acute, it is sometimes difficult to find out the reasons that guided their actions. Yet, having the ability to verbally articulate a reason for action must not be confused with having such a reason. It may be the case that somebody acts on a specific reason without being able to communicate this reason. Agents, e.g., who are temporarily suffering from aphasia due to a stroke, nevertheless display structures in their agency which refer to certain reasons for action. In babies not yet able to speak and in higher mammals, one can, from a certain point in the development of their cognitive abilities onwards, recognize reasons that guide their actions. Without the ascription of complex intentional states it is, however, impossible to talk of acting for reasons. In this case, the concept of action does not apply, and one would do better to speak only of 'behaviour'.

Acting expresses (accepted) reasons. To the extent to which behaviour expresses bare desires, it is, consequently, not guided by reasons and is thus not to be regarded as an action. Many actions, however, are – fully legitimately – directed at the satisfaction of various desires and needs of the agent. This is certainly not true of all actions, and it seems to me that the share of those actions which are solely directed at the satisfaction of personal desires and needs, is generally greatly overestimated. It is my impression that many – at least in our culture – believe that only those reasons can be regarded as expressing rationality which are related to the satisfaction of personal interests. This is the reason for the tendency to misconstrue one's reasons for action in a way that, *ex post*, relates them to one's own advantages. However, the thought that reasons are rational only when they are related to the promotion of the agent's well-being, is deeply misguided.

Inclinations never directly determine actions, but do so only via accepted reasons for action. Under certain conditions, the fact that an action satisfies a present desire is a good reason for choosing it. I have already mentioned some of these conditions; suffice it here to refer to intertemporal coherence and moral criteria such as considerateness for others. When we accept a reason for action, we also take a position with regard to what role the momentary desire should play in our agency. The fact that I have made a promise constitutes (under certain constraining conditions) a reason for action, and so does the fact that I have a specific desire. The fact that performing a certain action would be advantageous for me can be a good reason for performing it, in much the same way that duties are attached to the social roles and virtues that are relevant to the respective decision situations. Desires do not play any

special role in the large spectrum of elements that can be relevant for finding out about good reasons for action. As is the case with all other (empirical) characteristics of decision situations, they become relevant through the agent's *judgement*, in the form of the reasons accepted by him. Every action is determined by reasons. The distinction between moral reasons and those referring to self-interest is irrelevant here. Every action reveals the reasons approved by the agent. Freedom of action can only be accounted for when it is acknowledged that desires are never directly action-guiding, but determine actions only via reasons. Weakness of will is one form of losing this freedom. This is why the weak-willed person typically faces difficulties in justifying her actions. *A limine*, the behaviour of a weak-willed person cannot be characterized as agency.

The rational and non-akratic person chooses that action which seems most appropriate after weighing possibly conflicting *prima facie* reasons for action. This weighing takes into account subjective traits of the person such as his desires and interests. Nevertheless, the term "subjective reason for action" is an oxymoron like the notion of "subjective fact". A fact obtains, or it does not obtain. An action is justified, or it is not. For a person, however, a fact is never directly accessible but rather only by way of her beliefs; in other words, a person must be convinced that it is a fact that P. (The same is true of practical reasons.) This, however, does not render it a subjective fact that P. Rather, the person is convinced that it is a fact that P. A person who concludes that G is a good reason has a certain belief. One might call this belief "subjective". Yet, it is the nature of beliefs to be subjective. Much in the same way as facts, 'reasons' are 'objective'. A case of diverging beliefs about good reasons for action is to be dealt with analogously to a case of conflicting beliefs about matters of fact. Hence, reasons for action are objective, while beliefs about reasons for action are subjective.

Subjectivists with regard to the justification of actions, who are often proponents of the standard view, will object that this objectivist theory of reasons for action cannot account for the agent-relativity of reasons for action. If something is a good reason for P to perform h, it does not follow that it is also a good reason for somebody else to perform h. A theory implying that all persons have the same reasons for action would be inappropriate. This argument points out that a theory of reasons for action has to be sufficiently complex. Individual desires can be a good reason to do something, and, of course, not all persons have the same individual desires. If they have divergent desires, this implies that different actions can be justified for them. For one person, then, there is a good reason for performing a certain action due to the empirical condition that the person has an appropriate desire, while this reason for action does not obtain for the other person not meeting the same empirical condition (since he has, for instance, other desires). This form of agent-relativity is harmless and quite compatible with an objectivist theory of reasons for action.

Relativity should not be taken to be an indicator of subjectivity. Rather, relativity requires us to design our theory such that the relativity may be reconstructed within a framework of non-relative, objective criteria. In our ordinary discourse, a certain degree of agent-relativity of good reasons is accepted as perfectly normal, and philosophical theory should be able to account for this fact. Statements such as "for you

(but not for me), this is a good reason" can be interpreted in a subjectivist fashion: "to you, but not to me, this seems to be a good reason" (in which case this way of putting the matter rather than the first one should be used). It can also be interpreted in a relativistical fashion: good reasons for action vary from one person to another, and there are no 'person-independent' reasons for action. I take it, however, that such a statement should be interpreted in an objectivist fashion as follows: "Due to our common understanding of good reasons, this is, considering your present situation, an (objectively) good reason for you to do *a*, while I am in a different situation (e.g. since I have different desires) such that my doing *a* would not be good." As long as we seriously discuss reasons, we are implicitly committed to an objectivist interpretation of them. The way in which we think about and discuss practical questions is incompatible with a subjectivist or relativist interpretation.

Taking into account desires and interests as good reasons is compatible with an objectivist interpretation of practical reasons. When one considers the large spectrum of different kinds of reasons for action, the usual dichotomy between moral and non-moral reasons for actions seems less plausible.

In the first place, one can distinguish between reasons for action which only refer to consequences of actions and those which do not. The standard view is often tied to the additional assumption that every rational action is directed towards an optimization of its consequences. As long as the notion of the 'consequence of an action' is not extended such that all relevant traits of actions can be interpreted as a 'consequence', this *consequentialist theory of practical rationality* is false: there may be many good reasons for an action which are not exclusively tied to its consequences (regardless of whether they are causal or probabilistic consequences). When somebody has done me a favour, I have a good reason to be grateful to him and to express this gratefulness in my actions. This reason for action obtains independently of the consequences of the actions performed out of gratefulness. It may be that these actions lead to the other person's being inclined to do me favours in the future. This expectation, however, is not the reason for my being grateful, and even if I have reason to believe that the person will never be able to do me a favour again, I have a reason for being grateful (and to act accordingly). The reason for action, as it were, lies in the past: I am grateful (and act accordingly) since the person has done me a favour in the past. I do not have to know anything about the future in order to be grateful and to act accordingly. I might, of course, deliberate which action would please the other person, and to this extent I might need to deliberate about the consequences of my actions. This possibility does not, however, contradict the claim put forward before: my deliberations solely concern the question for an adequate choice among the plurality of possible actions which are expressions of my gratitude, while the genuine reason for my action – the fact that the person has done me a favour – remains in place.

Not only can the actions of others provide reasons for my actions, but also my own actions can provide reasons. When I have promised to do something, I generally have a good reason to keep this promise. This reason is independent of whether the consequences of my keeping the promise are optimal. If people were to only keep their promises in cases in which they expect to optimize the consequences of

their actions in so doing, the institution of promising would lose its important role of coordinating our actions. Yet, this does not imply that the institution of promising loses its coordinative role through individual acts of promise-breaking. Having given a promise generally constitutes a good reason for keeping the promise. This is another type of reason for action which does not refer solely to the respective action's consequences.

Also, the fact that somebody has requested me to do something, normally (and considered on its own) constitutes a good reason to do what I have been requested to do. Of course, there can be other, conflicting reasons which are weightier and which can outweigh the respective *prima facie* reason. This, however, leaves unaffected the fact that actions of others which were beneficial for me, a promise which I have given, or a wish that has been expressed, by themselves, are *prima facie* reasons to act accordingly: to show gratitude, to keep the promise, or to fulfil the wish.

Are reasons of the kinds just mentioned moral or non-moral reasons? It seems to me that in the face of the plurality of good reasons it would be artificial to distinguish sharply between reasons which promote my own well-being and which are therefore prudential reasons and those reasons promoting the well-being of others or which relate in some other way to the interests (or the rights etc.) of other persons and which are consequently counted as moral reasons. When I fulfil the wish of another person, I do so for a good reason but often without moral motivation. When I pursue my own interests, I have a good reason to do so only when the pursuit of my interests does not harm others.

The fact that good reasons for action often do not refer to the consequences of the action which they recommend speaks in favour of a structural theory of practical rationality and against a consequentialist reduction to intentions which are fulfilled by the consequences of the action. In most cases, the rational action is not the adequate means to cause certain intended consequences to occur in the world. The rational action represents accepted good reasons, some of which refer to favours done by others, some to given promises, others to wishes of other persons or to given orders, some to one's own interests or to the presumed interests of others, many to rights and duties, and several (directly) to ethical principles. This vast plurality of practical reasons is, as it were, the basic foundation upon which a theory of practical reason must be built, and it cannot be modified arbitrarily. Rather, every theory of rationality has to be corroborated with regard to this plurality. Since conflicts occur within this plurality of practical reasons, rationalization and ethical theorizing is called for in order to solve practical dilemmas and to provide orientation in cases in which life-world reasons fail. Integrating the life-world plurality of practical reasons and at the same time modifying life-world practical reasons by way of rationalization are not irreconcilable opposites, but rather work together as guidelines of good practical philosophy.

Cooperation

No society could exist without the willingness of its members to cooperate. There are so many different forms of cooperation that attempting to provide a general characterization of cooperation might seem a hopeless enterprise. Basic decision-theoretic techniques, however, allow for a definition of cooperation. A cooperative action can be defined as the choice of the dominated strategy in a situation of interaction which is of the same type as the prisoner's dilemma. This definition is in accord with the most important, if not with all of the usages of the term "cooperation" and this surprisingly simple characterization, therefore, is no arbitrary "nominal definition".

When a person cooperates with another, she, in a way, disregards her own interests, in order to make sure that something is achieved which both have an interest in. Cooperative actions are generally not altruistically motivated. The idea is not to do somebody else a favour and, in so doing, to put ones own interests last. Rather, two cooperating persons coordinate their actions such that a combination of actions occurs which is desired by both. In contrast to mere coordination, cooperating persons do not individually optimize their respective interests.

This difference between coordination and cooperation can be brought out using a simple example. Suppose two persons want to meet in a café, and there are two cafés to choose among in town. Without coordination on the café to meet in, there is the risk that one person goes to one café and the other to the other. Both have an interest in meeting in the same café, however. In order to act in a coordinated way, it is not necessary to disregard, in a way, ones own interests. Suppose, however, that two employees of a firm strive for the same (higher) position in the firm. If a person succeeds in showing the other's work in a bad light, the chances of the former to get the post rise. When both, despite of the competitive situation, refrain from talking bad about the other, a form of cooperation takes place. If none of the two were ready to engage in this restricted form of cooperation, they would talk bad about each other, respectively, and for both, the probability of getting the position would diminish. This is why mutual cooperation for both is better than mutual non-cooperation.

© Springer Nature Switzerland AG 2019

J. Nida-Rümelin, *Structural Rationality and Other Essays on Practical Reason*,
Theory and Decision Library A: 52,
https://doi.org/10.1007/978-3-319-95507-0_4

For the individual, however, the most advantageous case is one in which the other person cooperates while one's own actions are not cooperative. Whatever the other does, it is better for the individual to be uncooperative. Yet, when both are uncooperative, this is worse for each of them than is the case in which both cooperate.

One could ask: how can it be that both have an interest in cooperation while the individual does not have an interest in cooperation? This is possible since "both" can be used collectively: with regard to the comparison of the two situations in which either both cooperate or neither cooperates, there is a common interest in cooperation. From the individual point of view, however, cooperation is not advantageous. Cooperation is made possible when the individuals refrain from optimizing (be it optimization with regard to their individual interests or other values, including moral ones). This is why acting cooperatively is deemed irrational in traditional optimizing conceptions of practical rationality. For the conception of structural rationality, however, cooperation is a paradigmatic case of rational action.

The cooperating agent chooses his actions such that they are (potentially) part of an interpersonal structure of agency, which structure is preferred by the agent over the interpersonal structure of agency which would result from individually optimizing behaviour. In addition, the cooperating agent expects the other agents which are (potentially) involved in the cooperation to have the same preferences with regard to these alternatives, and he expects them to be equally willing to cooperate. Willingness to cooperate is exhibited in the fact that the agent contributes his share to mutual cooperation when he expects that the others will also contribute their share. One might suspect this way of putting the matter to involve a departure from methodological individualism. This, however, is not the case since the question of whether a person cooperates or not is answered solely with regard to properties of this person. An interpersonal structure of agency is cooperative if all agents involved cooperate and the reciprocal cooperative expectations are satisfied.

To act cooperatively is to chose one's action such that it is part of a collective action (= a combination of individual actions, = an interpersonal structure of agency), which collective action the person prefers to the collective action which would result from individual optimization and of which the person believes that the other agents (potentially) involved in the cooperation have the same preference. Willingness to cooperate is, in general, irrational when it is not expected that the others are ready to participate in the cooperation.

Cooperation is not a good reason for action under all circumstances. Cartels, i.e. intentional cooperation of firms, usually supported by unofficial arrangements, in order to weaken competition among the firms involved, are not desirable from the point of view of economy as a whole, since cartels involve overpricing and undermine the advantages of market competition. Some forms of cooperation in society conflict with basic norms of equality. Informal political circles of friendship in political parties, for instance, undermine the rules of democratic decisions in the respective institutions, which rules are laid down in the law for political parties and in the statues of the parties themselves. Party members not belonging to these informal groups cannot articulate their point of view in the same way as those who are involved. The very strong cooperative relations within families in Southern Europe

often lead, in form of nepotism, to unequal treatment and to a lack of transparency of public decisions. But if there are no external effects, no negative effects on those, who do not participate in the interaction, cooperation is (structurally) rational and morally justified, given that the preferences involved are acceptable.

Moral rules can at least partly be interpreted as a result of cooperation. It is in everybody interest that all follow the rule of keeping promises, that all respect individual rights, that all refrain from doing harm to others etc. Complying with them might be motivated cooperatively: I do my part, expecting that the others do their part in establishing structures of (collective) agency that are preferred by everbody given its alternative: a Hobbesian state of "competence, diffidence and glory" with the result that life is brutish and short, as he formulates it famously in *Leviathan.* Structural rationality does not require that you act on cooperative motives, but it renders it rationally possible.

The intention to choose one's own action such that it can be embedded into a cooperative structure of interaction is not necessarily irrational. A theory of practical rationality must not exclude intentions of this type as being irrational right from the start. Acting cooperatively can be irrational, i.e. conditions are possible under which the process of weighing reasons can result in a verdict against the choice of the cooperative action. In many cases, however, there are better reasons for cooperation. If this is the case and if it is rational to expect that the others are willing to cooperate (i.e. that they would cooperate if the others cooperate), then the transition from approving of a cooperative action to the choice of an individual cooperative action is paradigmatic for the conception of structural rationality. When a person has good reasons for approving of a structure of agency, then his individual action which is part of this structure is *eo ipso* rational. Our life-world normative arguments and our everyday practice support the basic intuition of structural rationality: *reasons for generic (structural) actions ipso facto reasons for those individual (punctual) actions which correspond to them adequately.*

If a situation of interaction is sufficiently symmetrical, i.e., if the situations of the relevant agents do not fundamentally differ in morally relevant respects, then only symmetrical interpersonal structures of agency can be the result of reasonable individual decisions. If there are no external effects to be taken into account, mutual cooperation is to be preferred over mutual defection (i.e. mutual optimization of individual interests). This preference for cooperation results solely from the individual preferences of the relevant agents. Further moral reasons do not have to be brought into play here. The standard view, however, does not allow for transferring individual preferences in favour of cooperative collective agency to the rationality of those individual actions which constitute cooperative collective agency. In other words it does not allow to justify punctual acts with their structural rôle. But that is what we do in lifeworld moral practice.

Life-world morality rests on a plurality of explicit and implicit rules for action. Probably all of them establish a dilemma of cooperation, independent of whether they themselves can be interpreted as solutions to a problem of cooperation. The individual agent can ask with regard to each established rule whether he should generally conform to this rule or not. Suppose the agent supports general conformity

to the respective rule on the basis of moral reasons; he supports, e.g., the rule to help the innocent needy. More precisely: he supports that in his society all or, at any rate, most persons who find themselves in a situation in which somebody is in need of help and innocently so, react in helping this person. Let us assume that every member of the relevant society approves of this rule. This being so, everybody, implicitly, adheres to a certain, let us say "humane" conception of his society. Individual conformity to this rule, however, is not relevant for the question of whether or not the society is humane in this sense. For each individual person, therefore, the moral reason which is decisive for him does not consequentialistically carry over to his own actions. He wishes a society in which a certain degree of conformity with the rule of helpfulness is upheld, but, by way of his own actions, can neither cause this to be so, nor can he cause it to happen that the actual degree of conformity falls short of the one desired by him. Seen from a consequentialist point of view, individual actions seem to be morally indifferent with regard to this rule. This paradox can be solved when the point of view of structural rationality is adopted: the moral reasons supporting general conformity to the rule of helpfulness are *eo ipso* good reasons for individual conformity to this rule.

A problem of cooperation can also occur with regard to moral preferences: every person prefers all (or sufficiently many) to follow the rule of helpfulness. Each individual person knows that the effect of his individual conformity with regard to the humaneness of the society can be neglected. The optimization of individual preferences would cause the humane structure of society not to be realized. Seen from the point of view of structural rationality, however, the individual person acts in accordance with the rule of helpfulness since he has a preference for a humane society and since he regards his own action as a part of the relevant interpersonal structure of agency.

Two different problems of cooperation occur with regard to those (moral) rules which can be interpreted as solutions to problems of cooperation: first, the problem for which the rule in question is a solution, and second, the problem arising due to the question of conformity. The rule can count as a solution to a problem of cooperation when collective conformity would guarantee an interpersonal structure of agency which all prefer over the one which would result from individually optimizing behaviour. For an adequate analysis, it is indispensable to distinguish between these two problems of cooperation, although their solution is the same, i.e. complete and general conformity to the rule. The social validity of any normative rule leads to a problem of cooperation. Not all normative rules, however, can themselves be reconstructed as solutions to problems of cooperation. The interaction between those seeking help and those helping is not a case of cooperation. Yet, with regard to rule-conformity – or, as one can also say, with regard to the social validity of the respective norm – there is a relation of cooperativeness between the (potentially) helping persons. The moral content of the norm is not determined by cooperation. The social validity, however, is so determined. It is one thing to determine moral obligation, and it is another thing to do what is morally called for. Conformity with regard to moral rules requires structurally rational agency, which sustains the required cooperation with regard to moral preferences.

When morality itself is determined via cooperation, it is more difficult to keep these two levels – the level of the constitution of morality and the level of practical conformity – apart. Let us assume that the inhabitants of the old part of a town have all come to the conclusion that it would be better if nobody drove in this part of the city with his private car. All have a preference for a state of affairs in which there is no private traffic in the old town. All have the possibility of parking their cars just outside the old town and to walk to their homes. So far, however, it is not legally forbidden to drive into the old town. If there were a vote on such a legal step, there would be unanimous agreement in favour of it. Given these preferences, there is no question about the common good with regard to the issue under discussion and with reference to the inhabitants of the old town. It is in the general interest to keep the old town free from private traffic (where "general" relates to the inhabitants of the old town). If the interests of others are not touched by this issue, there is a good reason for the normative rule "Do not drive in the old town with your private car". This norm can be justified to every inhabitant of the old town. The individual optimization of interests, however, would lead to continued private traffic in the old town. The overall situation is only marginally changed by an individual decision to enter the old town with a private car. Individual comfort, however, is increased considerably, and this by far compensates for the marginal change seen from the individual perspective. Again, the structure of interaction is of the type of a problem of cooperation as introduced above. Those not driving into the old town decide in favour of the cooperative action, and those driving into it optimize their individual interests. The overall optimization of individual interests leads to a state of affairs which is worse for all inhabitants of the old town than the state of general cooperation. Cooperation here refers solely to personal interests and to the individual preferences representing them.

With regard to the (*ex hypothesi* coinciding) moral preferences of the inhabitants of the old town, which comprise possibly full conformity with the rule not to drive in the old town with private cars, there is a second problem of cooperation: the person conforming to this rule chooses the action which is cooperative with regard to the moral preferences, the person not conforming decides not to cooperate with regard to these preferences. The morally cooperative person decides to contribute her share to bringing about the common moral goal. She does so independent of whether her individual contribution does indeed improve the situation with regard to the moral preferences. The decision not to drive in the old town with a private car is justified, since the corresponding general course of action is justified. And, indeed, a person counts as irrational with regard to our life-world understanding, when he puts forward reasons for a specific general course of action without contributing her share. This can be interpreted as a discrepancy between normative beliefs and actual individual actions. Of course, there might be convincing reasons to the effect that the prima facie-irrationality of such a discrepancy is rectified.

At any rate, a person leaving her car in the garage since she approves of a general course of action to this effect, can be rational. In order for this person to act rationally, it is by no means necessary to show that all those reasons which speak in favour of the general course of action (to sustain the quality of living in the respective

town) also speak in favour of the individual action. In general, at any rate, it is not the real causal influence which an individual decision has for the quality of living which motivates the person to leave the car in the garage, but rather the justified approval of the corresponding general course of action. One can dispute whether it is possible that an individual action cannot, while a collective action (or structure of agency) constituted by individual actions can, contribute to an improvement of the quality of living. Discussions of this sort, however, are academic in the negative sense of this term, since they do not account for the actual structure of motivation and the rôle of embedding punctual acts into structural agency for it, that is at the core of structural rationality.

Consequentialist theories claiming to get along without such a relation of embedding deeply conflict with our life-world reasons for action. The discussion revolving around the so-called prisoner's dilemma can be regarded as an impressive warning sign of this deficiency of consequentialist theories. Cooperation is a common practice. Cooperative reasons for action determine large parts of our behaviour. A theory excluding (genuine) cooperation as irrational cannot be adequate. But also, theories which can only explain cooperation or render cooperation rational in terms of complicated detours – dispositions, evolutionary stability or iteration – overlook an important feature of practical rationality, namely the justification of punctual intentions in terms of structural intentions. If somebody gives reasons speaking in favour of a general course of action, he also gives reasons speaking in favour of the individual action which realizes the general course of action (or would realize it provided that the others also cooperated as expected). The overall structural intention is directed towards a collective course of action, the punctual intention to an individual action. To rephrase this in more familiar terminology: "I intend to contribute my share, such that we act in this way. I do *a* (individual, punctual, concrete act) since ***a*** is my contribution to ***A*** (collective, structural, generic agency)."

Communication

Communicative rationality is a special form of structural rationality. I shall make this claim plausible in five steps:

1. Intentions provide linguistic entities with their meanings.
2. Language is a cooperative institution relying on and enabling cooperation.
3. The specific form of linguistic cooperation rests on certain basic communicative norms.
4. These basic communicative norms are not – in opposition to what is assumed in intentionalist semantics – conventional (i.e., solutions to problems of pure coordination), but enable cooperation (i.e. are solutions to problems of cooperation). This is why it is justified to juxtapose communicative and strategic rationality, especially when one is ready to concede that the attempt to trace back linguistic meaning to mental states by intentionalist semantics is superior to behaviouristic approaches.
5. The high degree of conformity of linguistic behaviour to the basic communicative norms is an expression of structurally rational agency.

Ad (1). Intentions provide linguistic entities (and other, non-linguistic signals) with their meanings.

No meaning without intention. If bees do not have intentions it is only metaphorical to talk of a "bee language". To move up and down at a certain angle, then, does not *mean* that "there is food in this direction", but rather *causes* behaviour in other bees which is successful with respect to finding food. Yet, should the bee moving up and down have the intention to inform the observing bee about the fact that there is food in the respective direction, and should the observing bee ascribe this intention to the bee moving up and down due to her moving up and down, which leads the observing bee under favourable circumstances to the correct belief that there is food over there, then the moving up and down has the *meaning* to indicate the direction in which there is food to be found (the favourable circumstances referred to would

© Springer Nature Switzerland AG 2019

J. Nida-Rümelin, *Structural Rationality and Other Essays on Practical Reason*,
Theory and Decision Library A: 52,
https://doi.org/10.1007/978-3-319-95507-0_5

be a good faculty of judgement of the observing bee and truthfulness on the part of the bee moving up and down). The respective movement (the 'signal' or 'sign', as we could also say) has this meaning independent of whether there is a *convention* among the bees which assigns this meaning to this sign. For the Gricean model a speaker, with an utterance of his, *means that p*, if he intends to cause the hearer to believe that *p*, by recognizing that the speaker, with his utterance, has had this intention. But there is a vast multitude of intentions that are specific for certain situations corresponding to a multitude of different types of speech acts. Despite of this vast variety, it is plausible to assume that the *conventional meaning* of a sign depends on the fact that speakers using this sign have the respective *situational intentions*. Without speaker's intention no meaning.

Ad (2). *Language is a cooperative institution relying on and enabling cooperation.*

When somebody says something to somebody else, he normally does not do so without a reason, i.e. he can give an answer to the question why he has said whatever it is he has said. Different types of (illocutionary) speech acts open up a spectrum of possible answers: he wanted to inform, to warn, to command, etc. His intention can be further characterized with reference to the utterances. In doing so, however, the intention guiding the respective utterances does not become a linguistic entity. Beings without a language also have beliefs and act on intentions. Linguistic meaning can be reduced to non-linguistic facts – beliefs, suppositions, intentions, desires, etc.: an utterance has a (situational) meaning when it is used to realize specific intentions of the speaker by way of seeing to it that these intentions are recognized by the hearer. It has a general ('conventional') meaning when there are situations of a certain type in which it is generally used for realizing intentions of a certain type by way of making sure that these intentions are recognized by the listeners on the basis of the utterance. Communication as it actually takes place in specific situations is logically prior to (conventional) meaning: I do not understand the (conventional) meaning of a linguistic entity, unless I understand communication in specific situations.

There is a general interest in communication: each person prefers living in a society whose members are able to communicate with each other. This does not imply that all individual communicative acts are desired, and even if a specific piece of communication is desired, its content can run counter to the preferences of the hearer. Diverging interests with regard to individual acts of communication, however, do not affect the fact that there is a general interest in communication. Language is a cooperative institution. It is an *institution* since it relies on (normative) rules which are accepted by all (basic communicative norms), and it is a *cooperative* institution since it requires a high degree of conformity to its constitutive rules, which would not be achievable by individual punctual optimization. Punctual optimizing agents (with their optimization referring to their own interest or to other values) would in too many cases not follow the constitutive rules and would, ultimately, render communication impossible.

Ad (3). *The specific form of linguistic cooperation rests on certain basic communicative norms.*

With regard to utterances intended to transfer information, it is clear that successful communication depends on the truthfulness of the speakers, i.e. on the fact that the speakers only say things which they believe to be true: "only claim something if you yourself believe it." If members of a linguistic community do not conform to this norm of *truthfulness* and if this is known to the hearers, then the speaker's utterances are not of informative value, i.e. the listeners mistrust them.

The basic communicative norm of *reliability* needs to be distinguished from the norm of truthfulness. Members of a linguistic community should only claim something to be the case if they have good reasons for their respective belief. This second basic communicative norm is not a version of the first one since many are inclined to adopt beliefs without appropriate reasons. The basic communicative norm of truthfulness ensures a correspondence between claim and subjective belief. This correspondence, however, does not suffice for informative communication to be successful. The reason is that this type of communication does not only refer to subjective states of beliefs of the members of a linguistic community, but also to the "outer objective world". Since this world is not directly accessible and since we can hope that convictions are more likely to be true when they are based on good reasons, truthfulness is not sufficient.

The basic communicative norm of reliability does not require one to be philosophically ascetic in adopting beliefs ("adopt a belief only when you are able to justify it"), but to be ascetic in one's utterances: "claim something to be the case only if you can back up your claim with good reasons." This norm does not commit one to discretion. In order to clarify this, one could add: "In case you do not have good reasons for a belief, you can communicate the belief; you should not, however, claim it to be true. Beliefs not backed up by reasons should be communicated by way of a report about what one believes, and they should not be presented as a claim about what is the case in the world." Good reasons invoked here are good epistemic reasons, i.e. reasons to believe something to be true. Sometimes, there are non-epistemic, practical reasons for claiming that something is the case, even though one is convinced that it is actually not the case. The basic communicative norm of reliability can be interpreted as a principle of responsibility with regard to the other members of the linguistic community. With my (uttered) claims, I influence the beliefs of others, and I am obliged to contribute to the formation of beliefs that correspond to the facts as closely as possible.

This last remark points to a philosophical problem that is highly relevant for the conception proposed here. From the point of view of structural rationality, there is no clear-cut dichotomy between *epistemic* (or theoretical) reasons, which are involved in truth-tracking endeavours, and *practical* reasons, which refer to personal inclinations and aims of the respective persons (see Ch. 1 and 3). Practical reasons are expressed by (normative) beliefs. In the case of a special type of linguistic actions (claiming something to be the case), both types of beliefs and their respective guiding reasons are relevant. These two types of beliefs and reasons

cannot be integrated punctual but only within the framework of a comprehensive ethics of communication. In the last resort, the weighing of normative reasons is decisive for a linguistic action. The role of communication for social cooperation as a whole and the danger that informative linguistic acts are rendered impossible due to a lack of conformity with the basic norms referred to above have to be taken into account. In general, norms of truthfulness and of reliability will be given moral priority over other normative aspects.

The basic communicative norm of reliability renders the presence of (epistemic) reasons the adequate normative criterion. It could be objected that these reasons are not directly accessible and that the basic norm of reliability, therefore, needs to be put in a subjective way: "claim something to be the case only if you *believe* that you can back up your claim with good reasons." This norm, however, would be too weak. Reasons always refer to beliefs of the person having these reasons. Consequently, the norm requiring that one only utter those claims which are backed up by beliefs that are, in turn, guided by reasons, is not particularly strong. (Epistemic) reasons are relative to the respective system of (descriptive) beliefs and conjectures. To further weaken this basic communicative norm would favour the credulous in an unnecessary way. The basic communicative norms do not require truth but reasonable truthfulness.

A person can be truthful without being reliable. The converse, however, does not necessarily hold: a person cannot be reliable without being truthful, if it can be assumed that the person's beliefs are being formed on the basis of reasons (there are, of course, counterexamples). It is possible that a person's beliefs be formed on the basis of reasons even though he adopts some beliefs for which he has no reasons. For persons adopting those and only those beliefs for which they have good reasons, the basic communicative norm of reliability requires the same linguistic behaviour as does the basic communicative norm of truthfulness. Since persons of this kind, however, hardly exist in reality, the basic communicative norm of truthfulness is generally insufficient, and the basic communicative norm of reliability must also be postulated. Due to the possibility of the anomalous way of forming beliefs mentioned above, it is useful to maintain both norms, even though conformity to the basic communicative norm of reliability in most cases implies conformity to the basic communicative norm of truthfulness.

The third basic communicative norm is trust. When rational persons have reason to believe that the members of their linguistic community conform to the two basic communicative norms of truthfulness and reliability to a high degree, they trust in the correctness of each other's claims. To trust the claims of other members of the linguistic community means to take the others to be truthful and reliable. The asymmetric relation can be brought out by noting that the basic communicative norm of trust does not have a normative status when conformity with the other two basic norms can be assumed. It is in one's own punctual interest to trust the claims of a truthful and reliable person.

The question arises whether there are other norms, beyond these three basic communicative norms, which are *constitutive* for a language (in the sense of a system of communication). At any rate, if the argument presented above points in the right

direction, it is obvious that communicative agency in the weak sense of Habermas, i.e. communication about facts and so-called agent-relative practical reasons, which have no claim to universal validity, is only possible on a strong normative basis. It seems to me that intentionalist semantics and discourse ethics contain similar mistakes when it comes to grounding the normativity of communication. Reference to intersubjectively shared standards of evaluation does not render language a cooperative institution: an institution that presupposes the respective normative bonds. This leaves unaffected the fact that the critique of a purely 'strategic rationality' in the sense of intentionalist semantics is convincing in light of the reasons laid out above. No form of communication is merely a conventional coordination of interests. Rather, communication is interaction, the legitimacy and effectiveness of which vanish without a normative framework supporting the cooperative institution of language.

The cooperative institution of language requires conformity to the basic communicative norms mentioned above and, additionally, to the specific linguistic conventions which render the usage of the respective linguist means for communication successful. The requisite high degree of conformity to these specific rules of linguistic use, which are further characterized in speech act theory, can itself be regarded as a consequence of the conformity to the basic communicative norms, such that an additional normative basis does not seem necessary (this is true regardless of whether the speech acts are illocutions or perlocutions). It is obvious that linguistic means can be used not only for communication, but also 'strategically'. Using language strategically, however, can only be successful as long as there is sufficient conformity to the basic communicative norms, which back up the cooperative character of language.

Discourse ethics maintains that, over and above the basic aforementioned norms, there are further norms which are necessary conditions for discourse (i.e. for a form of communication aimed at enlightenment and understanding) and so – at least in the version put forward by Karl-Otto Apel – claims to provide a foundational justification for a specific ethical theory. Leaving its problematic metaphysical premises and the one-sided emphasis on the *homo disputans* aside, the rules proposed by discourse ethics are solely those norms which support the cooperative system of language (in the sense of a system of communication). A high degree of conformity of verbal behaviour to these rules is not a transcendental condition of communication, but nevertheless an important element of a comprehensive normative ethics, which includes linguistic interaction according to certain criteria. The rules proposed by discourse ethics, however, do not in themselves provide a sufficient foundation for a theory of normative ethics. There is no ultimate justification for them and they have a merely supplementary character in contrast to the three basic communicative norms discussed above. In the weaker version proposed by Habermas, discourse ethics can be characterized by the following principle: norms are valid if and only if they are (or can be) consented to in a practical discourse by all those concerned. Under the assumption that the discourse is neither personally constrained nor restricted with regard to the content of what is discussed (everyone can participate, and anything can be said or questioned), a norm will only be consented

to in practical discourse if "the consequences and side effects of its general observance for the interests of each individual in every particular situation can be accepted by all."

This raises the question as to what it means that a norm can be consented to by a person. Pareto-inefficient norms, i.e. norms which make at least one person less well-off than he could be (without making anybody else worse-off), would have no chance of being generally consented to. Among the Pareto-efficient norms, however, some norms are more just than others. Discourse ethics provides no criterion for settling issues of justice. The conflict of interests can only be solved by invoking substantial ethical considerations and not by solely referring to the individual interests and individual readiness to introduce them in a domination free and rational discourse. Cooperation in the sense of structural rationality requires a normative point of view that goes beyond securing the prerequisites of communication.

Reasonable persons wish to live in a social environment in which the rules proposed by discourse ethics are accepted and complied with. This is why structurally rational agents will comply with them in a punctual fashion as well (even in situations in which this conformity runs counter the optimization of individual interests), if they expect that the other members of the linguistic community will also comply with these rules to a high degree. Discourse ethics is a special case of structural rationality. The rules of discourse ethics are socially established in a society of structurally rational agents.

Ad (4). Since the basic communicative norms of truthfulness, of reliability and of trust are not conventional – in contrast to what is assumed in intentionalist semantics – i.e. they are not solutions to problems of pure coordination, but rather constitutive of cooperation (i.e. solutions to problems of cooperation), it is justified to contrast communicative and strategic rationality, especially when one grants that intentionalist semantics' way of explaining linguistic meaning by recurring mental states is superior to behaviourism.

The linguistic (and other) means we use for communication are conventional to a high degree: we could communicate using other linguistic (and other) means, and we conform to certain (established) rules for communication since it is better (even if possibly not optimal) to communicate using these means rather than making communication difficult or even impossible due to a lack of commonly accepted rules for communication.

A convention *C* can be characterized through the following conditions:

(1) Every member of the linguistic community conforms to *C*.
(2) Everybody believes that (1).
(3) Due to (2), everybody has a good reason to conform to *C*.
(4) If almost all conform to *C*, then everybody prefers that everybody conforms to *C*.
(5) There are alternatives to *C*: there is at least one other possible convention *C*', for which (3) and (4) would equally hold.

In this sense, many of the rules which guide our linguistic behaviour are conventional, i.e. governed by convention. In this respect, the basic communicative norms, such as truthfulness, trust and reliability, are different: one should not regard them as conventions but rather as solutions of generalized (iterated) situations of cooperation. Additionally, one can, at least with regard to specific types of speech acts, ask to what extent the general conformity to linguistic conventions (more precisely: to those rules which are constitutive for the respective speech acts) does not itself require a *normative* explanation.

The classical texts on intentionalist semantics by Grice, Lewis, Bennett and others, contain, at a central stage of their arguments, a fallacy which is analogous to the fallacy of inferring act-utilitarianism from rule-utilitarianism. When general compliance with a rule maximizes overall utility, this does not imply that every instance of compliance to this rule maximizes overall utility. Analogously: if there is a general interest in communication – i.e. if communication is in every individual's interest – and if communication requires general compliance with certain rules, then this does not imply that every individual member of the linguistic community has an interest in complying with these rules in all situations. Rather, there is a well-known phenomenon – much discussed in ethical theory – which is known under as 'free-riding', i.e. a way of acting which is not in conformity to the respective rules, but which is especially advantageous for the individual when the others continue to comply with the rules. With regard to rules which are solutions of situations of cooperation, typically that person profits who conforms less than the others. Punctual optimizing behaviour would start a spiral of non-conformity, which would ultimately undermine the basis of communication. Communicative rationality is cooperative, and the communicatively rational person complies (punctual) with the rules constitutive for communication, even if this is not in her (punctual) individual interest.

The definition of a 'convention' given above (a definition that followed David Lewis' proposal) is inadequate. The reason is that it leaves unclarified the difference between mere coordination and cooperation, and thus also between rules which are merely conventional – which, once established, lead to general conformity of all individuals involved, since this is in each individual's interest – and rules of cooperation, which cannot be sustained unless (communicative) acting guided by structural rationality and not by the individual's punctual interest in optimization. The conditions (3) and (4) in the definition of a convention given above are not valid for rules of cooperation (and, generally, for rules securing structural rationality), when a practical reason is understood in the widespread way which I have criticized in chapter "Acting on Beliefs", i.e. as expressing a momentary and self-oriented desire. Under this interpretation of a practical reason, a person is especially prone not to conform to the rule in a specific instance, if he can expect the others equally not to conform. Hence, the improbable case is precluded that a person's punctual non-conformity in every individual case would result in (or, at any rate, make probable) general non-conformity and, herewith, the decay of the rules securing communication. The preference for punctual non-conformity is independent of whether and to

what extent others conform to the rules. In cases of a low degree of conformity, however, talk of established rules of communication is no longer adequate.

Rational agents can successfully employ the means for communication only when they can rely on common rules and on common knowledge associated with these rules. A naïve 'rational choice' conception committed to the idea of punctual optimization could not make plausible the requisite high degree of rule conformity. Punctual optimization is only possible within the constraints set down by the rules constitutive for communication. Acting with regard to mutual understanding or communication and (punctual) optimizing (strategic) action are interrelated in the individual action. Actions that depend on a process of communication are bound by those rules which sustain the system of communication. In many cases, the respective rule-conformity is rendered 'automatic' by way of our learned speech behaviour to such a degree, that we are not consciously aware of this. The choices of actions take place within the constraints laid down by the rules constitutive for communication. With regard to the rules, however, 'parasitic' behaviour is possible by utilizing the rule-conformity of other participants in communication in order to deviate from these rules, if this leads to the attainment of other desired goals. The permanence of a linguistic community would not be possible without a high degree of conformity to the basic communicative norms of truthfulness, reliability and trust, and this requires structurally rational agency.

Structural Intentions

A person can have two basic kinds of propositional attitudes: *epistemic* and *conative* attitudes. A *propositional attitude* is an attitude of a person towards a proposition.

If I expect the sun to shine tomorrow, then I have a certain *epistemic* attitude towards the proposition that the sun will shine tomorrow. The proposition is that the sun will shine tomorrow and my epistemic attitude towards this proposition is the expectation, or: my subjective probability, which is close to or equal to 100%, that this proposition is true. Epistemic attitudes, in the form of (descriptive) beliefs, expectations, conjectures, certainties, hypotheses, etc., concern what is, has been, or will be the case. If a person's epistemic attitudes are coherent, then they can be represented by a subjective probability function.

If I wish that the sun will shine tomorrow, then I have a certain *conative* attitude towards the proposition that the sun will shine tomorrow. The proposition is that the sun will shine tomorrow and the conative attitude towards this proposition is the one of wishing. Conative attitudes, in the form of wishes, hopes, intentions, etc., concern what is, has been, or will be the case. If the conative attitudes of a person are coherent, then they can be represented by a subjective desirability function.

Attitudes can change over time, but their character is *dispositional*: they are manifested in respective judgements and actions, provided that certain conditions are satisfied. To this extent, the epistemic state of a person is an abstraction: what is meant is the totality of the epistemic attitudes of the person at a specific point in time. Since the epistemic attitudes of a person are not isolated from each other, but rather are interrelated in complex ways and comprise specific observational judgements as well as theories and background beliefs, one may speak of an epistemic system. The same is true for conative propositional attitudes. Since both basic types of propositional attitudes are interrelated with regard to their manifestations (judgements and actions), the system of epistemic and conative propositional attitudes cannot be separated in two distinct parts.

Changes in a rational person's propositional attitudes do not occur arbitrarily, they are not even caused, but are guided by reasons. This is true both of epistemic

© Springer Nature Switzerland AG 2019

J. Nida-Rümelin, *Structural Rationality and Other Essays on Practical Reason*,
Theory and Decision Library A: 52,
https://doi.org/10.1007/978-3-319-95507-0_6

and of conative attitudes. Ultimately, the rationality of a person displays itself in the structures – including the temporal structures, i.e. the dynamics – of his epistemic and conative attitudes and the agency resulting from them.

Actions are affected by intentions of different sorts. There is no action without intentionality. Three types of intentions can be distinguished, a distinction that does not cut across different types of contents of these intentions: (a) intentions which are constitutive for the actions which they, so to speak, accompany (agency as controlled behaviour); (b) intentions which precede the actions and which are satisfied by the action itself (decisions); (c) intentions which motivate the choice of an action (motivating intentions) and which are not satisfied by the action itself but, e.g., by the action's consequences. Motivating intentions do not have to be satisfied by the consequences of the action, since the embedding of an action in larger structural contexts can also satisfy motivating intentions. In this chapter, the intentional side of structural rationality will be examined.

Intentions are a specific sort of conative attitudes, and they can be distinguished from other conative attitudes by their intimate connection to actions. They are not simply given, but, rather, they are the result of a person's weighing theoretical and practical reasons. They are embedded in the system of epistemic and conative attitudes as a whole.

If one assumes a separation between epistemic and conative attitudes (a separation which ultimately cannot be upheld) for the purpose of the further argument, then the following is the case: a great many of the epistemic states of a rational person are the result of a weighing of (theoretical) reasons. Many of the conative state of a rational person are the result of weighing (practical) reasons. Beliefs that are direct (causal) consequences of observations are not under the control of a process of weighing (theoretical) reasons. One can grant the existence of such beliefs without having to abandon the coherentist interpretation of the weighing of reasons, since beliefs of the type just mentioned are too marginal to be able to carry all of the justificatory weight for our system of beliefs. In a similar way, desires are not subject to a weighing of (practical) reasons. (Descriptive) beliefs are sometimes highly dependent on theories, e.g. on beliefs about the number of the stars in the universe or the causes for the phenomenon that the sky appears to be blue. Our life-world (descriptive) beliefs are largely independent of theoretical assumptions. We can agree on life-world facts without necessarily having to agree on (scientific) theories. We can communicate with people in a life-world way who do not have any theoretical scientific knowledge. The conative attitudes which are guiding our life-world agency are largely independent of ethical theories. We can agree on reasons and counter-reasons without having to agree on ethical theories; and others lacking any knowledge about ethical theories can criticize our actions.

Not only epistemic but also conative propositional attitudes are shaped by reasons. There are more or less basic beliefs and conative attitudes, which serve as starting points of theoretical and practical justificatory chains. There is no general criterion that is able to discriminate between basic and not basic elements. It is not the logical form (the form of quantification), not the level of generality or concreteness, it is the rôle they play within our lifeworld reasoning that defines their

epistemic status. There are quite concrete 'observations', i.e. epistemic beliefs that are in large degree (but not solely) dependent on sensual perception that are to be count as basic in the epistemic sense, generic assumptions like the spatio-temporal constitution of physical objects, but also e.g. moral rules of non-discrimation or the positive value of a concrete act of helping a person in need of help. Conative attitudes and practical reasons are closely interrelated, in much the same way that epistemic beliefs are linked to theoretical reasons. Some of our beliefs are not subjected to further justification and reasonable critique, and similarly some of our conative attitudes are excluded from justifiability and critique with reference to practical reasons. But since the class of propositional attitutes that cannot be subjected to further justification varies over time, culture and context, we cannot adopt a foundationalist point of view. Everything can in principle be doubted if we realize that there are inconsistencies and incoherences within our system of propositional attitudes. No set of normative or descriptive propositions can be fixed in advance (in advance of reasoning) that is beyond justifyability but at the same time without doubt. But only if such a set existed, epistemological foundationalism would be justified.

Actions are not causal consequences of conative attitudes, directed by descriptive beliefs as the standard view assumes. Actions are the result of a normative judgement which integrates theoretical and practical reasons. Actions do not directly unveil the mental states of their agents. The privileged access of the acting person to her own mental states remains intact. To this extent, logical behaviourism as a philosophy of language is misguided. An action is not solely rendered rational as a means to realize something given (descriptive belief) and something irrational (conative attitude). The rationality of an action is an expression of the rationality of the epistemic as well as the conative attitudes.

In order to justify a descriptive belief, we fall back on other descriptive beliefs that are not doubted, and to those connections between beliefs which are equally beyond doubt (assuming regularities, laws, theories). Theoretical reasons explicitly refer to other propositions which seem to be certain – often they explicitly refer to theoretical (although not necessarily scientific) assumptions that relate these (more) certain propositions to the (more) uncertain ones; this happens by the way of inclusion in a whole system of background assumptions, a system which cannot be conceived as a whole and which includes general descriptive assumptions – for example on the structure of space and time – ways of theory-building, conceptual prerequisites, rules of inference, etc.

The situation is perfectly analogous in the realm of practical reasons. The justification of a conative attitude includes uncontested conative attitudes and those connections (regularities, normative rules, normative theories) which relate uncontested elements of our conative system to each other and additionally allows them to be related to the contested elements. Formally, practical reasons can be regarded as the markers of those conative attitudes which are responsible for the fact that these conative elements are action-guiding for a rational person. Hence, an enormous system of implicit background beliefs is utilized: a system which, e.g., comprises certain assumptions of invariance (for instance conditions of fairness), criteria of

relevance, methods of normative theory-building, rules of deontic logic, but also basic normative principles.

The rationality of a descriptive belief depends on the way in which it is embedded into the structure of the epistemic system. The rationality of a conative attitude depends on its embeddedness into the structure of the conative system. With regard to descriptive beliefs, philosophy of science has provided many insights into the different relations of embeddedness. Some justifications are ultimately of a deductive structure, others are inductive; the most important relations in science, however, are reductive or abductive – this is sometimes called the "inference to the best explanation". These different forms of arguments represent different ways of embedding, they do not stand in opposition to each other, but rather form the framework of justification in our life-world as well as in the sciences.

Practical justifications differ from theoretical ones with regard to what is being justified, but not in their form of justification. Deductive, inductive and reductive arguments also play a role in practical justifications. There are punctually and structurally descriptive beliefs, beliefs which refer to singular states of affairs and beliefs which refer to regularities and laws, and similarly there are punctual and structural normative judgements. Our descriptive beliefs can be, at least ideally, represented in a coherent probability function. Our normative judgements can be, ideally, represented in a coherent function of desirability. This is why decision theory does not oppose a differentiated analysis of epistemic and conative attitudes, but is rather a method of testing elementary standards of coherence and a way of uniformly (reductively) representing conative and epistemic attitudes. In decision theory the representation of epistemic and conative attitudes by probability and utility (desirability) functions, does not allow for disentagling them. This interrelation has a natural interpretation in the interwoven structure of conative and epistemic attitudes. In this chapter, we will only consider a specific aspect, namely, justification by embedding into intentional structures.

Let us start with an especially innocent example. I am sitting in a restaurant, and of all the available dishes, I would most prefer having a plate of *spaghetti alla marinara*. The desire to have *spaghetti alla marinara* is motivated by the expectation that I will have certain pleasant sensations, especially when this dish is combined with the open white wine offered here. The expectation of a pleasant sensation is a good reason to form the wish (the conative attitude) to have a plate of spaghetti alla marinara now. This good *prima facie* reason can, of course, be outweighed by other conflicting reasons, such as the fact that I have seen, in passing, that the kitchen is dirty or that the spaghetti has been cooked a day in advance. In the process of weighing the reasons, these latter reasons would have a greater weight, such that the original desire to have *spaghetti alla marinara* would not persist. Presumably, the desire in question can be satisfied by actions on my part. This desire is the result of a weighing of (*prima facie*) reasons for action. When I order a plate of *spaghetti alla marinara*, this action is motivated by the intention to be able to perform later the action of eating the *spaghetti alla marinara*. For ordering, I need to perform a number of other actions (I may have to do something in order to get the waiter's attention, to say certain things, to turn around, etc.). I perform all of these actions not for

their own sake, but in order to place my order; and I am placing my order not for its own sake, but in order to be able to eat a plate of spaghetti. These 'in order to'-relations are relations between intentions.

When I intend to perform a specific action *a*, and when I (descriptively) believe that performing *a* requires the performance of actions $a_1,\ldots,a_n$,, then an overall (structural) intention is formed, which comprises the total sequence $<a_1,\ldots,a_n>$. In this case, $a_1,\ldots,a_n$ would be a sequence of actions that are instrumentally chosen in view of action *a*; *a* is intended, and $a_1,\ldots,a_n$ are intended solely by virtue of *a*.

Even if there is more than one way to perform *a*, then the desire to perform *a* is (under certain circumstances) a reason for a 'way' *A* or $<a_1,\ldots,a_n>$. Those who try to maintain an optimization conception of practical rationality can interpret this feature of our life-world practical justifications in the following way: since the time of the deliberation about alternative options (ways) is limited, it does not make sense (it would lead, as it were, to further costs) to require a further deliberation about the specific choice of *A* as opposed to the other actions that would realize *a*. Due to a lack of sufficient information, the person is consequently indifferent to these options of which she is at least partially aware. Therefore, the fact that one of these ways *A* leads to *a* is sufficient to justify *A*. However, there is a different conception of rationality which makes no reference to the notion of optimization According to this conception, the justified desire to carry out *a* is a good reason to pick *A* even if there are alternative ways *A*', *A*",…. An additional element of justification (beyond the fact that *A* is a way to *a*) for *A* is required only when there are *prima facie* reasons speaking against *A*. Buridan's ass dies of starvation because he is indifferent to two stacks of hay and does not find any reason for choosing one of them over the other. When indifferent, the optimizer chooses one of the optimal actions, yet he is unable to give a reason for his choice. The structurally rational agent can give a reason for his way of acting. Our life-world understanding of practical rationality requires our conative attitudes to be coherent. Our intentions must satisfy certain criteria in their relations to one another. Our life-world understanding of practical rationality is at odds with punctual optimization.

Intentions are realized in actions that have various temporal and personal extensions. My intention to stop smoking tomorrow ranges over the remainder of my life-time. The intention to carry a piano in another room together with three other people is realized in an action that extends over four agents and a relatively short period of time. Realizing an intention in an action is not completely under the control of the acting person. Between the decision and the action there can be intervening circumstances, and in the case of intentions that are realized by multi-agent actions, there is a dependency on the decisions of the other agents involved. There are intentions which are more comprehensive, and those which are less comprehensive, depending on whether they are realized in more or less extended actions.

More interesting than the temporal and the personal aspect of the extension of an action is its decomposition into sub-actions. Few actions are basic in the sense that they cannot be decomposed into other actions. Accordingly, the properties 'punctual' and 'structural' are relative (and gradual): an action *a* is punctual with regard to another action *A* insofar as *A* is realized by the performance of *a* and further

actions *a*', *a*", etc. In order to perform the action of ordering a plate of *spaghetti alla marinara*, a plurality of individual actions must be carried out. These individual actions a_1, a_2, …, a_n are punctual with regard to the action *A* of ordering a plate of *spaghetti alla marinara*. It is possible that raising one's arm in order to call the waiter is a basic action, i.e., under normal circumstances this action is not part of the performance of other, more punctual actions. Having said this, however, I do take the idea of a repertoire of basic actions to be misleading. Whether or not an action is a basic action depends on the context and especially on the intentional state of the acting person. In the end, actions are individuated by intentions.

If *A* is performed by way of carrying out a_1, a_2, …, a_n, then a_1, a_2, …, a_n are justified in virtue of *A*'s being justified. If I desire to order a plate of *spaghetti alla marinara*, and if I can do this by carrying out a sequence of actions such as 'turning round in my chair', 'raising my arm', 'uttering certain sentences', 'nodding in response to a query', etc., then every punctual action in this sequence of actions is justified in virtue of the fact that the action of ordering is justified (by my desire to eat a plate of *spaghetti alla marinara*). The action of ordering is structural with regard to the individual actions, the performance of which realizes the action of ordering. *My structural intention justifies my punctual intentions*. Most will agree to this with regard to the example of ordering a plate of spaghetti. The justificatory relations between structural and punctual intentions can be taken to be almost trivial in cases in which the performance of the structural action is sufficiently 'close' to the performance of the punctual actions. The larger the range, however, the greater are the conflicts between this triviality, on the one hand side, and certain, especially consequentialist, views about rationality, on the other.

My punctual action to do my share in carrying the piano is justified by a structural action constituted by the collective carrying of the piano (by four persons). If there are good reasons for the respective structural intention, then *ipso facto* there is a good reason for me to contribute to this activity. It is, of course, possible that a person has this structural intention, and also has reason to question whether she should participate. In such cases, there is a practical conflict between opposing *prima facie* reasons. The person may, for example, take it that a refusal to contribute to this activity would lead to a situation in which there is another person contributing her share, such that the desired result (the piano is in the other room) could be attained without physical effort on the part of the deliberating person. If this reason dominates in the deliberation, then the structural intention to carry the piano into the other room together with three other persons is no longer present, and hence the punctual activity of one's own contribution can no longer be justified with reference to the structural intention.

Justificatory relations are not solely from top to bottom, i.e. from structural to punctual intentions. Someone who intends to stop smoking has to take into account the disadvantages that go along with not lighting the next cigarette, especially in the early stages after having stopped smoking and in this way having realized some important structural intention. The person has coherent intentions and is, in this respect, rational only when the structural intention is stable with regard to information about the desirability of punctual actions which realize (punctually) the

structural intention. The formation of rational intention requires both, the ability to embed punctual intentions into structural ones and to evaluate structural intentions regarding given or expected punctual ones.

There are different ways in which a theory of practical rationality can try to avoid the structural aspects of rationality. One possibility is the attempt to interpret the structural intentions such that they logically imply punctual intentions. The structural intention to stop smoking tomorrow implies the punctual intention not to smoke tomorrow or (even more punctual) not to light the first cigarette after having eaten breakfast. Slightly modifying our characterization of structural intentions, however, renders this way of construing the matter inadequate: let the structural intention be to stop smoking as soon as possible. When this is assumed, the argument – known to many, if not all smokers who have already tried to quit smoking – that it is actually irrelevant whether one stops smoking today or tomorrow (after having smoked for 20 years) develops its full force. And, indeed, my state of health in 5 years from now might be completely unaffected by one further day of smoking. Since the actual goal of my structural intention to stop smoking as soon as possible is to have (5 years from now) the characteristics of a person who has never smoked before, this argument – as claimed by medical studies – would even be correct. Applying this argument 1825 times, however, will lead to a situation in which my state of health in 5 years from now is one of a person who has smoked for 25 years.

In order to make this analysis fully convincing, one could supplement it by invoking the idea of preference thresholds. If the punctual view of agency is such that with regard to the consequences relevant for performing the structural intentions, the effects of punctual actions are under certain thresholds, then there is no possibility to transfer the motivation for the structural intention to the lower level of the punctual intention. A punctually optimizing person would undercut the entire complex of his – as we shall assume – justified structural intentions, i.e., he would punctually act in such a way that none of his structural intentions would be realized (or only accidentally so).

A theory of practical reason tries to clarify the motivations for actions of an ideal (rational) person. It is, however, a decisive characteristic of rational persons that they chose their actions in view of their conformity to structural intentions – entirely independent of whether there is an alternative way of calculation which does not take into account the idea of structural embedding. It is rational to choose punctual intentions with regard to structural ones and to choose a punctual action with regard to a sequence of actions that realizes a structural intention.

Weakness of Will

A person acts out of *strength of will* if she punctually does what she has realized to be structurally rational. *Weakness of will* is displayed in structurally irrational punctual optimization. Strength of will is a prerequisite for responsible action.

There are many motives which may prevent a person from conforming to the structures he himself has subscribed to. Momentary inclinations lead to ways of action that do not conform to what seems sensible on the long run. Everybody is inclined to discount the future to some degree: i.e., reasons referring to temporary distant effects are typically taken less seriously than reasons referring to more imminent effects of an action. Viewed from a long-term perspective, this leads to (intrapersonal) structures of agency which are not desirable from the perspective of the acting person. The problem does not only concern the weighing of short-term vs. long-term consequences of actions. It often occurs that an action is advantageous when assessed from a punctual perspective which takes into account its long-term consequences, even though the intrapersonal structure of agency resulting from punctual acting of this sort is disadvantageous.

The interpersonal problem of cooperation (see chapter "Communication") has an intrapersonal analogue. The analogy results from the fact that persons have different inclinations and assessments at different points of time. Metaphorically speaking, different conative and epistemic attitudes occurring at different points in time have to be coordinated in order to make a coherent life possible. Interpersonal variances have their analogue in the variances occurring within a person over the course of his lifetime – both with regard to his conative and his epistemic attitudes.

The conceptual problems which the phenomenon of weakness of will (*akrasia*) poses for the theory of practical rationality are, in the first place, an indication of the fact that the standard conception of practical rationality is defective in a fundamental way. The reason is that rationality, under this conception, is instrumental with regard to giving epistemic and conative states. The phenomenon of *akrasia*, therefore, has to be reconstructed within the framework of given conative and epistemic states at a certain point in time. It is obvious that this is at odds with the life-world

© Springer Nature Switzerland AG 2019

J. Nida-Rümelin, *Structural Rationality and Other Essays on Practical Reason*,
Theory and Decision Library A: 52,
https://doi.org/10.1007/978-3-319-95507-0_7

experience of weakness of will. Weak-willed action is experienced as a conflict between good reasons on the one hand side, and giving way to momentary inclinations on the other.

I have argued in this Essay that actions are never the product of inclinations, moderated by epistemic attitudes, alone, but represent accepted practical reasons. Within our view of structural rationality, therefore, the problem of weakness of will (*akrasia*) can be reformulated as follows: How is it possible for a person to accept a reason for action and act on this reason, although she herself believes that this reason should not be decisive, but that it is in fact outweighed by other (structurally more comprehensive) reasons for action?

Interestingly enough, the insight that *akrasia* results from a conflict between punctual optimization and structural rationality is the most plausible interpretation of the respective passages in the *Nicomachean Ethics*. The fact that these passages are normally interpreted in a different way indicates that the standard view cannot account for *akrasia* adequately. Aristotle asks what it is that the akratic person acts against. "The one opinion is universal, the other is concerned with the particular facts, and here we come to something within the sphere of perception; when a single opinion results from the two, the soul must in one type of case affirm the conclusion, while in the case of opinions concerned with production it must immediately act (e.g. if everything sweet ought to be tasted, and this is sweet, in the sense of being one of the particular sweet things, the man who can act and is not restrained must at the same time act accordingly). When, then, the universal opinion is present in us restraining us from tasting, and there is also the opinion that everything sweet is pleasant, and that this is sweet (now this is the opinion that is active), and when appetite happens to be present in us, the one opinion bids us to avoid the object, but appetite leads us towards it (for it can move each of our bodily parts); so it turns out that a man behaves incontinently (acratic) under the influence (in a sense) of reason and opinion, and of opinion not contrary in itself, but only incidentally – for the appetite is contrary not the opinion – to right reason. It also follows that this is the reason why the lower animals are not incontinent, viz. because they have no universal beliefs but only imagination and memory of particulars."

The interpretation of *akrasia* that Aristotle gives deviates from our structural interpretation of *akrasia* regarding only one, but interesting, point: For Aristotle the conflict is not between two normative beliefs, two reasons for action, but between an inclination causing the action and a normative belief, based on good reasons, telling against this action. For the account of structural rationality this would be only the marginal case in which the agent looses all control over what he does. In the normal case of weakness of will, however, acting acratically is still a kind of acting, whereas this marginal case could not count as acting any more. To give a concrete example, the drug addict feeling himself forced to take the drug immediately, becomes that, what Frankfurt called a Wanton, a being that does whatever its desires at a given point of time tell, not modified by deliberation. A Wanton cannot be held responsible for what it does. Strictly speaking a Wanton doesn't act, not even acratically. Aristotle is right that beasts not having the capacity to act on reasons do not act acratically. Beasts behave as Wantons. are. In this case acting on

punctual reasons is also acting on reasons, the difference is but between a normative (structural) attitude

An overweight person who, following a momentary appetite, eats more than he himself deems reasonable, does not act without a reason. To interpret this action as solely expressing a momentary inclination would be an inappropriate analysis. There are overweight individuals who choose their dishes deliberately, who base their decisions on the expected quality of the meal, and who even take into account information such as calorie content. One cannot demote, so to speak, the resulting action to a mere reflex to a momentary inclination. In fact, the overweight individual can give a conclusive answer to the question why he made this decision. In his answer, taste, conduciveness to good health, and inclinations will play a role. The choice of the dish is an expression of accepted reasons for action, although this choice is, *ex hypothesi*, structurally irrational. In many cases, the person would be inadequately described if we were to say that the degree of his understanding of the situation varies over time. Normally, it is not the case that the overweight person realizes after a meal that there have been important reasons speaking against eating this meal. Rather, he was aware of this already when ordering the course and while eating, or, at any rate, it was possible all the time to make him realize this, and this realization would not have influenced his way of acting. It is an inadequate analysis to say of the weak-willed that they follow not reasons, but inclinations.

The revealed preference-conception in decision theory which says that actions represent preferences (or, more precisely, that one ought to ascribe those preferences to a person which she reveals through her actions, given appropriately offered alternatives) can be transferred to the conception of structural rationality and to its picture of agency. Actions are interpreted as an expression of the weighing of reasons only if the acting person is considered completely rational. The weak-willed person, however, is not completely rational. She acts on reasons although she knows that there are stronger reasons speaking against this way of acting. She does not, however, act without any reason whatsoever, also the weak-willed agent is responsible for his actions.

Is it true that we do not hold young children responsible for what they do? It is true in the juridical sense, and we do not resent their wrong doing. But we ask them quite early in their life, why they did something. Responsibility is a gradual concept, too. Children learn to act responsibly in experiencing that they are held responsible and fail to act accordingly. To hold them responsible means to ask them which reasons guided their behavior. Children begin to give answers to questions of this kind roughly at the same age they begin speaking. They name the goal they wish to attain at the beginning of this phase, often only by pointing to some object, which signals that *this* is the object of their desire. The intentions they pursue are often short-term, and it is easy to stop young children from pursuing their goals by redirecting their intentions to something else. Their behaviour, so to speak, depends on the moment. They react to changes in their environment by abandoning their previous goals or by forgetting them; they quickly despair when they do not achieve something which they desire to attain, and their moods can quickly change. The behaviour of little children is weak-willed, it lacks the structuring of long-term reasons for action, but

it is not completely irrational. Also little children act, they are not Wantons and they learn that it is better for themselves not to be Wantons. Good part of education is about this: to learn that it is better for oneself and others not to behave like a Wanton. This insight stands in sharp contrast to the fact that mainstream theories of rationality take the Wanton to be perfectly rational.

In order to ascribe some behaviour to a person as one of his actions, it is necessary that we take her to have had reasons on which she acted. The small child who points when queried, refers to a reason for why it acted as it did. Hence, it is clear that this behaviour qualifies as an action. The guiding reason, however, is short-term and specifically related to the situation. It results from a momentary inclination, a desire arising in the child to immediately satisfy a wish. For the child, this desire is irresistible in the strongest sense of the word; accordingly strong is the despair if this desire cannot be satisfied. Although it is easy to redirect the child's intention to some other object, hereby leading him to abandon the previous goal, it is impossible for the child to dissociate itself – e.g. guided by the insight of how difficult it would be to attain the goal – from a present desire. Metaphorically speaking, the distance between the occurrence of a desire and the orientation of one's behaviour towards desire-satisfaction is quite small. As an intermediate element, there is only a very rudimentary estimation of the appropriate means to satisfy the desire. The distance is too small to allow for more complex instrumental 'weighings of reasons'. In order to strive for some goal by way of entering detours, distancing one's self from the momentary inclination would be necessary, in addition to the ability to maintain desires in the face of experiences that show how difficult it is to satisfy them. The young child is a punctual and short-sighted agent and, consequently, is not able to realize most of its goals, the young child lacks structural rationality. The personality of the child grows with its ability to control larger parts of its behaviour and thus to be more and more responsible for its own life. Larger and larger parts of its behaviour qualify as agency, and the person will be more and more visible in the lasting structures of her agency.

The weak-willed person also acts following a normative judgement, yet a judgement which reasons do not consider the long-term consequences, the week willed person disregards structural reasons. Weakness of will is expressed by the fact that a person inappropriately weighs the momentary traits of the situation she is in. What she takes to be counting in the situation is what is temporarily closer, although this runs counter her better judgement. The person can correctly weigh the relevant reasons for action, deciding, however, for an action which does not (completely) conform to the result of this weighing. Nevertheless, the action of the weak-willed person is expressing a normative judgement – one that shows what the person takes to be important in this very moment.

This can be easily clarified with reference to the example of an addicted person. An addicted person willing to be treated for drug addiction often acts against his own better judgement. He knows that – seen from a structural point of view – it would be best to stop the addictive way of action. Nevertheless, he decides in the individual situation in favour of continuing the addictive behaviour, since he inappropriately overweighs the punctual traits of the situation. He thus avoids the

negative consequences that would be foreseeable, at least in the first phase of a cure for drug addiction. Since these consequences are not actually present, but only expected at the moment of the decision, it becomes clear that addictive behaviour itself must generally be counted as agency. It is the result of a[n] – albeit too punctual – weighing of reasons. The addicted person *decides* for the short-sighted avoidance of negative consequences over the long-term goal of no longer being addicted. The irrationality of this way of acting lies in the fact that this punctual optimization does not accord with what is structurally desired. In his behaviour, the addict is similar to a child: punctual optimization dominates, whereas a long-term structural way of acting on accepted structural reasons does not take place. In a minimal sense, the addicted subject remains responsible for his actions, his weakness of will, however, diminishes the degree of his responsibility. Only in the extreme case, in which the addict already feels the momentarily negative consequences of stopping the addictive behaviour and loses the control over his behaviour completely he could not be held responsible any more.

Weakness of will does not pose any conceptual problems for structural rationality. In acting, we follow reasons. We do not, however, always act according to our better insight into the relevant reasons. A deviation from the best judgement is called weakness of will, and this can be reconstructed as a conflict between structural reasons on the one hand, and punctual reasons on the other. To put the matter positively: strength of will is displayed in structurally rational agency. Responsibility for individual behaviour increases with the ability to structure rationality, i.e. to integrate individual actions in the larger structures of one's agency.

Structural Coherence

Behaviour qualifies as action if it is controlled by a person, i.e., by her intentions and these intentions are an expression of her (normative) judgement. Behaviour increasingly becomes agency as it is accompanied by rationalization. It is, however, not very attractive to characterize a reasonable person by saying that she deliberates all the time upon available actions, that she permanently controls her own behaviour, that her practical life is characterized by an intense sequence of normative judgements. One way of dealing with this problem is to rely on the claim (which can be formalized in decision theory) that the person should deliberate upon different options for actions only in those situations for which the expected costs of this weighing (time is money) are lower than the expected utility that results from the fact that decisions based on weighing reasons are in many cases better than those lacking such a basis. The conception of structural rationality, however, points in a different direction: a reasonable person is not characterized by constraints on her punctual deliberations, but by the fact that she embeds her actions into larger structures of agency.

Punctual optimization requires to take many, to many decisions. Structural decisions for an overarching rule (generally: structure) of agency reduce the number of decisions that have to be taken. The choice of structures is an expression of individual freedom which provides the conditions for ethical habitualization. Individual and collective forms of life represent these structures. A virtue (aretè) is a habit (hexis), but also a decision (prohairèsis) says Aristotle in the Nicomachean Ethics, the second characterization is mostly overlooked by present day aristoteleans. Structural decisions establish patterns of behaviour that make deliberation in the singular concrete superfluous. In the ideal case the remaining deliberation concerns exclusively the question which feasible actions fit best into the structures chosen. While the punctual optimizer is solely constrained by the epistemic limits of his deliberation, the structurally rational person knows of the superiority of structured behaviour; she does not allow punctual optimization to destroy structures, and thus she does not, in comparison with the punctual optimizer, reach the limits of her

© Springer Nature Switzerland AG 2019

J. Nida-Rümelin, *Structural Rationality and Other Essays on Practical Reason*,
Theory and Decision Library A: 52,
https://doi.org/10.1007/978-3-319-95507-0_8

deliberative capacities as quickly. In contrast, an increase in practical rationality can reduce the complexity of deliberation, thus allowing the subject to concentrate on essential matters. *A limine*, the structurally rational person chooses a form of life, she remains free in every singular instance to deviate from the structures constitutive for this form of life, and she nevertheless chooses her actions such that they support these structures and keep them in place. The virtual figure of the structurally rational person takes only one singular decision for a coherent form of life, and herewith renders unnecessary all punctual deliberations.

In reality people go back and forth, i.e., they choose certain structures and they modify these structures when confronted with specific decision situations. Real people know about the uncertainties of their basic existential decisions, and they give them periods of probation; they confront them with specific experiences they encounter when putting the existential decisions in place, and they specify, modify, and, if necessary, revise these basic decisions. To this extent, for real people the basic existential decision does not stand at the beginning in an isolated way, but rather is in a constant tension to everyday individual decisions. The punctual optimizer knows only causal constraints, whereas the structurally rational person remains free and nevertheless enjoys the advantages of practical reason, namely the possibility of leading a coherent life.

There are no external criteria for the correct choice of a form of life. The conception of structural rationality does not opt for a new normative foundationalism. The normative weighing within structural rationality cannot rely on indubitable principles of evaluation, and does not provide a theory determining the intrinsically valuable, independent of the specific forms of life. We do not have, in contrast to classical utilitarianism, a hedonist measure of evaluation at our disposal that assesses forms of life solely with regard to the question to which integrals of subjective well-being these would imply. The good form of life results from the balance of good normative reasons. Among these reasons are not only moral reasons in the narrow sense, e.g., reasons which refer to cooperation with and considerateness to others, but also reasons taking into account individual inclinations and desires. Ultimately, the necessity to provide justifications only occurs in cases in which there is incoherence. Internal incoherence requires a justification from the acting person himself, while external incoherence requires justification in the face of critical questions of others. A good justification resolves the incoherence which has given rise to the problem or the question – intrapersonally or interpersonally. Decisions against which there are no good *prima facie* reasons do not require further justification. The analogy with beliefs is obvious. Reasonable doubt – i.e., critical questions worth answering – crop up as good *prima facie* reasons speaking against the belief in question. Other kinds of scepticism do not have any place in the life-world (*Lebenswelt*): they would be 'academic' in the worst sense.

Structural rationality is the practical analogue to the subsumption of theories in the natural sciences. Embedding individual theories into more comprehensive ones, such that individual theories can be characterized as special cases of the latter, is an advance in rationality even when no further empirical phenomena can thereby be explained. By way of such an embedding, the coherence of the respective complex

of descriptive beliefs is increased. Whether the more comprehensive theory itself is justified, is determined in part by the fact that it successfully allows for the embedding of corroborated 'smaller' theories. The justificatory relations do not only start from the overarching theory and end in their specializations, but also vice versa. The fact that a corroborated theory can be embedded speaks in favour of the more comprehensive theory. The same is true for the relation between more comprehensive structural decisions (in the extreme case, existential decisions) and 'smaller' and 'more punctual' decisions. A more comprehensive structural decision is corroborated when 'smaller' decisions, for which there are good reasons, can be embedded into the larger structure. On the other hand, the smaller decisions are more weighty and plausible if they can be embedded into a structural framework.

The analogy between the coherence of descriptive beliefs and the coherence of decisions is far-reaching. In both cases, the justificatory relations are coherentist and holistic. There are no deductive, axiomatic structures of justification, and there is no sharply separated realm of the indubitable, and thus there is no reason to adopt the standpoint of normative or epistemic foundationalism. In both cases, formal frameworks or theories are available: decision theory and deontic logic in the practical and classical logic and the theory of probability in the theoretical sphere. Decisions represent normative beliefs. Normative beliefs concern questions of right acting. Normative judgements are judgements as all others, they obey (with little modifications) the same logic, and they should be interpreted in a realist (or, better, objectivist) way. To this extent, the analogy claimed above is indeed a triviality. Decisions represent conative attitudes or normative beliefs of a very specific type, and the justificatory relations between them are analogous to the ones between epistemic attitudes (or descriptive beliefs).

The analogy between normative beliefs which guide actions and descriptive beliefs can, however, be misleading, when one considers scientific hypotheses and theories as paradigmatic cases of descriptive beliefs. Instead normative judgements should be seen as analogous to descriptive everyday judgements and not to scientific hypotheses. Normative judgements are, different from scientific theories and hypotheses, a constitutive part of the Lebenswelt (life-world) and therefore more resistant to revision, they are deeply interconnected with our everyday patterns of agency. There are no revolutions regarding our form of life, although there are changes over time, but these evolve and need time, different from science, where genuine revolutions are possible. This important difference does not affect the fact that justification of beliefs of any type is coherentist and this includes the justification of decisions, since decisions represent a specific type of normative beliefs, as we have seen.

Long-term (structural) decisions can justify small (punctual) ones. The long-term decision to undertake a certain journey justifies numerous smaller decisions such as calling a travel agent, ordering a ticket, etc. Under certain circumstances, big decisions constitute good reasons for action. The question, "Why are you doing that?", when asked in reference to a short-term decision, can oftentimes only be sufficiently answered by appealing to a long-term decision. Long-term decisions can constitute reasons for action. This asymmetric relation between long- and

short-term decisions does not, however, contradict the coherentist understanding of structural rationality. This becomes clear when we further consider the alternative possibility of a deductivist and fundamentalist understanding of structural rationality. Whether a decision is long- or short-term is a relative matter. The decision to travel to Brazil is long-term when compared to the decision to call the travel agent in order to buy the ticket. The decision to travel to Brazil is short-term in comparison to the decision to go on a longer holiday every year. There is a hierarchy of levels of decisions, which, in the end, amounts to the choice of a form of life: the choice of a life as the ultimate axiomatic instance of justification, so to speak, for individual actions and decisions. But where could we find the criteria that justify a reasonable choice of the form of life? Either there must exist external criteria of justification, the resources of which are beyond the life-world structures of justification, or else the conception of structural rationality would suddenly terminate as decisionism. Both of these options are unattractive.

External criteria for evaluation are themselves in need of justification. Such justifications cannot be obtained without reference to our life-world justificatory practice. All large rationalist attempts of justification of this type can be regarded as having failed. This is true of Jeremy Bentham's justification of a utilitarian principle, as well as of R. M. Hare's attempt to derive an ethical criterion from the logic of moral discourse, or of the alleged ultimate justification of norms as a transcendental condition for communication. Now, one could hope that it might be true that those normative criteria which regulate interpersonal relationships cannot be justified without the appeal to given moral structures of justification, but that this is not true of criteria for assessing individual forms of life. This objection, however, is highly implausible: if it is not even possible for relations of interaction to develop ethical principles with axiomatic validity, then the assumption that such an attempt is worthwhile with regard to the plurality of individual forms of life is most implausible.

The decisionistic interpretation of structural rationality would not allow for the existence of good reasons for the choice of a form of life. This choice would be, so to speak, the ultimate 'existentialist' act of self-determination or of choosing a life. This radical form of existentialism, however, cannot be reconciled with the fact that we have reasons for organizing our life in a specific way and that these reasons refer to a plurality of specific conditions for decision. We justify our decisions by appealing to respect for others or cooperation, and we devote part of our life to projects, the realization of which we believe to be intrinsically valuable (independent of our interests), we organize our lives such that we experience satisfaction and enrichment. A reasonable person makes large (or 'long-term') decisions – and, if possible, the biggest (or most permanent) one, i.e. the decision for a form of life, while keeping in mind the specific situations of life to which the envisaged decision will probably lead. He makes big decisions not only with regard to their embedding into even bigger decisions, but also in view of the smaller decisions determined (justified) by it. A big decision, to this extent, is justified with respect to the small decisions which are, in turn, justified by the big decision itself. This involves no circularity but

exhibits the coherentist character of practical justifications. The conception of structural rationality can only be made sense of within a coherentist framework.

When I have made a good long-term decision, then *ipso facto* those small decisions are justified which are necessary in order to put the big decision into practice. In many cases, it is unclear which of the possible small decisions are best suited to realize the big decision, and different big decisions sometimes conflict in light of specific small decision situations. In the process of weighing the priority, one can resort to the next higher level and thus increase the coherence of the structures of decisions, or, in the individual case, there is an arbitrary resolution of the overdetermination. Big decisions are corroborated in those small decisions which are justified by the former. When a big decision necessitates a small decision which is unacceptable for ethical reasons, then the adequacy of the big decision is called into question. The relations of justification between decisions are consequently deductive as well as reductive.

A form of life which is completely coherent in itself does not pose any internal problems of justification. This does not imply that it might be unacceptable for ethical reasons. For the acting person, however, there are no justificatory questions within the framework of a fully coherent form of life. Analogously, problems of justification in the interpersonal case only occur when the *prima facie* justifications for decisions of the involved persons diverge. Interpersonal coherence, however, does not mean identity of forms of life. Two persons, for instance, who agree that different inclinations, different cultural backgrounds and different social roles create different conditions, in view of which different decisions can in turn be justified, are not facing a mutual problem of justification.

In accepting a reason for action, one commits oneself to a certain structure of one's agency. A person who justifies one of his actions by saying that he wanted to be considerate to somebody else or, better, that only this action meets the criteria of considerateness, hereby commits himself to also act considerately in other situations which are similar in relevant respects. Of course, a person is neither causally nor logically, but rather ethically compelled to do this. It suffices to accept that our life-world normative justifications become practically relevant in this way, and an analysis of the most important ethical theories shows that this life-world fact is taken into account in one way or another in ethical theory formation. Reasons correspond to structures of agency; the acceptance of reasons corresponds to the respective big decisions. How can this observation be made compatible with the coherentist understanding of structural rationality? Is it not such that the acceptance of normative reasons ultimately ties one to a top-down procedure? The first part of the answer might be surprising: we accept normative reasons only to the extent to which they provide intuitively plausible guidance for action in specific situations of decision. Normative reasons, consequently, are subject to corroboration, and they are not simply given. The second part of the answer refers to the specific framing of the content of the normative reasons. Being considerate to the interests of others is doubtless a good reason for action in certain situations. Yet, it would be hopeless to determine an abstract criterion of considerateness which is equally adequate for any decision situation. It only shows in the specific situations of action what it means to

be considerate to other persons. The abstract reason for action of considerateness is, so to speak, interpreted in the specific situation of decision. In our punctual agency, we show how we interpret this general reason for action and the weight we assign to it. Crucially involved in this is, for example, the ability to empathize, i.e., the ability to adopt the perspective of another person in a specific situation. Only when confronted with real situations of decision, do general criteria for agency become sufficiently specific. The everyday use we make of the general reason for action determines its meaning. This use has to be sufficiently coherent, meaning that the corresponding structures of agency have to be supported with a uniform normative justification.

This far-reaching parallel between the coherentist justification of decisions and actions on the one hand, and descriptive judgements on the other, however, should not lead one to overlook one important difference. Obviously, decisions play a different role in practical reasons than in theoretical ones. When I have decided to go to a certain restaurant (I could have decided to go to another one), then I have a reason to turn right into this road or to call this restaurant in order to reserve a table. The decision renders certain other decisions rational: the decision constitutes good reasons for action. In the case of theoretical reasons, decisions do not play a similarly important role. Surely, the choice of a system of coordinates, for example, is relevant to the description of natural states of affairs; this, however, does not affect the justificatory status of the respective descriptive judgements. In philosophy of science, it is largely that background beliefs play a crucial role in theory formation. A shift from one paradigm to another leads to different forms of descriptions and does not allow – provided that Thomas S. Kuhn is right – for the direct comparison of theories. The status of a paradigm with regard to scientific justifications is similar to the status of existential decisions with regard to life-world practical justifications – insofar as the justification of hypotheses takes place within a paradigm and the justification of small decisions takes place within an existential decision (a form of life) – and they are not (and cannot) be brought into play from an external point of view. This does not contradict the fact that there are, e.g., ethical reasons which cannot be disposed of in an individual form of life. Each form of life has to satisfy ethical principles which it integrates into its structures of agency, and in so doing it makes them a crucial component of their justification.

Decisions play a more important role for practical justifications than for theoretical ones. How this relation can be described in the case of the existential decision on the one hand, and paradigm shifts on the other is in need of a more specific analysis. We prepare this analysis by first looking at the case of ‘small’, not ‘existential’ decisions and their role in practical justifications.

For most of our small decisions, we can give reasons. When a decision fits especially well in the established structures of agency and the corresponding relations of justification, it will not be doubted. If, however there are alternative options for decisions and if the choice of one of them is not immediately appealing, then the expectation arises that it is justified. The following cases can be distinguished.

(1) The case in which the options are of equal value if all relevant information is considered. If a person believes that both options are of equal value, she is indifferent to them. In order to avoid starving like buridan's ass, the person will just have to make up her mind in order to resolve the situation. In doing this, she does not decide to believe that one option is of higher value, but rather sticks to her initial assessment of the equality of the values of the options. The decision is justified in a double sense: it is justified by the idea to choose the best of all possible options and, secondly, to resolve the situation in which there are two or more best options.

(2) If the fact that a person cannot decide between two options or does not recognize sufficient reasons for preferring one option over the other is due to incomplete knowledge, then the situation is to be distinguished from the case just considered. Two sub-cases need to be considered:

(2a) The first sub-case is one in which the expected values of both options are equal in the face of given probabilities. Consequently, the person is rationally indifferent to the options in the decision-theoretic sense. In a world without objective probabilities, this situation, however, would be different from case (1), which is characterized by incomplete knowledge. If the existence of objective probabilities can be assumed, case (2a) can be subsumed under the case in which the options are of equal value which can be known in the case of complete knowledge.

(2b) The second sub-case is one in which no probabilities can be assumed. The consequences of the decisions are, accordingly, unpredictable, which leads to a specific form of helplessness that should not be called indifference.

(3) The fact that none of the possible decisions can be preferred over another may be due to a case of conflicting reasons.

(3a) The reasons speaking in favour of one and the other option could lead to the fact that both options are regarded as being of equal value. (3a) should, however, not be confounded with (1). (3a) is to be understood to the effect that conflicting reasons occur, i.e., reasons which speak for one and the other decision, respectively. A paradigmatic case of (1) is a case in which a person pursues a goal and this goal can be achieved by two decisions which are, in this respect, of equal value. In case (3a), there is, in contrast, a case of a genuine conflict. In deciding to go for one of the possible actions, the person decides against considering at least one other reason for action. It is assumed here that the person can weigh the reasons against each other; that he comes to the conclusion that they are of equal value and that it is due to this fact that both options are of equal value. In most cases of conflicting but comparable reasons, however, the practical weighing leads to a determinate result and characterizes one of the open options as being supported by weightier reasons.

(3b) From the situation just described, we need to distinguish the sub-case in which the conflicting reasons are incomparable or incompatible. The person can be thrown into a helplessness of a completely different kind with regard to the question which of the options he should go for. Rich material on situations of this kind is provided in the literature on moral dilemmas.

(4) The most extreme form of practical helplessness occurs when no action-guiding reasons at all are available in a situation of decision. Whereas in case (3) there is an overdetermination of the situation, we now have a case of underdetermination. The totality of practical reasons does not cover the whole realm of decision situations. In the case under discussion, there is a situation in which no reason for action appears to be applicable. The interpretation of "applicable" is crucial here. It is, indeed, hard to imagine that it should be impossible that among the vast complex of practical reasons there is not a single one applicable to the respective situation. It is enough to assume that the possibly applicable reasons are not sufficient to be considered as action-guiding. Situations of this kind often occur when the conditions of human agency are changed, e.g., by the emergence of new technology. The options which have become, or are likely to become, possible due to advances in genetic engineering provide numerous examples. The present boom of applied ethics in different fields can be regarded as a reaction to this specific kind of helplessness.

In the cases (1), (2a) and (2b) and (3a), there is no incoherence in the system of practical reasons. At most, one could suspect an incoherence in the system of theoretical reasons in (2b) since one might claim that it should always be possible to assign probabilities to incertain alternatives. Equality of value of the sort (1), (2a) and (3a) is, however, no reason to suspect problems of coherence. It is just important that equality of value does not lead to the inability to decide, i.e., that the person makes up his mind, whatever the decision is.

Even when in these cases there is no unresolved conflict of practical reasons for an incomplete normative assessment, the decision has a specific status. It is not merely the expression of a successfully performed process of weighing or an expression of the acceptance of overwhelmingly obvious reasons without any deliberation. Rather, it is arbitrary if one of these cases occurs. The result of the decision could have been different, although there are no convincing reasons in favour of the alternatives. Decisions of this sort are, in a specific sense, 'shaping'. The indifference resulting from the process of deliberation must not seduce one into believing that it is irrelevant whether a person decides in this way or in another. In the sciences, too, there are cases which are characterized by an equilibrium of theoretical reasons speaking in favour of one and of another hypothesis. In cases like this, one can also speak of a certain arbitrariness of decision. Nevertheless, this situation is different, since such a choice of hypotheses is always made with a reservation. As soon as the equilibrium of theoretical reasons is resolved by additional information, the chosen hypothesis can be subjected to revision. This is different in cases of small life-world decisions which, once they are made, influence the further course of things and which have, in themselves, a justificatory force.

We face a serious problem of coherence in the cases (3b) and (4), i.e., in dilemmatic situations of decision or in situations in which no reasons are applicable. The incompleteness or incompatibility of the given system of practical reasons forces the person to decide in order to resolve the problem. This decision can, in itself, have justificatory force since it resolves an incoherence or incompleteness of a

decision that is not justifiable within the given system of reasons. This is the difference between dilemmatic (3b) and underdetermined (4) situations on the one hand and situations that are characterized by equality of value as in case (1), (2a), (2b) and (3a).

When dilemmatic or underdetermined situations occur with regard to small decisions, this raises questions that transcend the specific situation and its possible resolution. The accepted reasons have to be revised. Since accepted reasons determine a form of life, it may be necessary, in order to achieve coherence, to partly modify these reasons.

While dilemmatic situations and underdetermination rarely occur in small situations of decision which are embedded into a form of life, they seem to be characteristic for big (existential) decisions. This is so because existential decisions involve a shift in normative standards. The decision itself changes, so to speak, the criteria for its own assessment. If we take 'existential' decisions to be decisions that are 'big' enough to deeply influence a form of life as a whole, then it is obvious that standards of rationality for small everyday decisions are no longer in place. Even when a large part of the reasons for action remains intact when one moves from small to big decisions, since these reasons are invariant with regard to the forms of life – and this is especially true with regard to some of the ethical reasons – surely most reasons for action depend on the chosen form of life. Existential decisions such as one's career choice or one's choice of a partner – can therefore only be rationalized to a lower degree than is the case with small decisions. The coherentist embedding of a decision into the context of a form of life and the guiding reasons can be applied to existential decisions only in a very limited way and, in extreme cases, not at all. The prudential maxim to weigh the reasons carefully when it comes to unimportant decisions and to follow intuitions in the case of really important decisions, therefore, has a rationale. If weighing our reasons for actions that are embedded into the respective form of life does not provide any, or only incoherent, guidance, then a decision that reduces incertainty is reasonable. In existential situations fearful hesitation and repeated deliberation mostly does not lead to more practical rationality, but rather threatens one's steadfast pursuit of the chosen path. The person's identity might even be threatened, if she is unable to fix what she takes to be her authentic way of life. Deliberation has its limits. The feeling of being at peace with oneself is brought about by a structurally coherent form of life; it is part of a good life and helps to live a good life.

Part II: Other Essays on Practical Reason

Why Consequentialism Fails[1]

Introduction

The paradigm of consequentialism is an ethical theory: utilitarianism. The traditional critique of (act-)utilitarianism confronts some practical implications of this theory with our moral intuitions. I think that this traditional critique is indeed quite successful. It is successful because normative theories cannot be justified *more geometrico*: a good normative theory can be developed out of one principle alone, but the justification of this principle is the successful role of this principle to make our normative judgements coherent. If some principle is unable to do this, then it fails, even if it may have appealing qualities in terms of explicitness, simplicity and universality. Even if the traditional critique of utilitarianism and – more generally – ethical consequentialism is successful, it has one major weakness: it overlooks the fact that ethical consequentialism can be understood as a specification of consequentialism (with regard to actions) in general. To put it in more explicit terms; If there were good reasons to adopt a consequentialist theory of rational action, ethical consequentialism – even if not necessarily in the form of classical utilitarianism – is a natural consequence. The specific strengths of ethical consequentialism are based on the forcefulness of a consequentialist theory of action.

In the critical literature on contemporary utilitarianism this close linkage between the theory of rational action and utilitarianism has mostly been neglected. The reason is partly that most theorists criticizing utilitarianism are either deontologists, who quite seldom take rational-choice theories to be adequate, or (anti-normative)

Published in "Why Consequentialism Fails", in G. Hilström-Hintikka and R. Tuomela eds, Contemporary Action Theory, Vol II., Kluwer, Dordrecht, 1997, pp. 295–308.

© Springer Nature Switzerland AG 2019

J. Nida-Rümelin, *Structural Rationality and Other Essays on Practical Reason*,
Theory and Decision Library A: 52,
https://doi.org/10.1007/978-3-319-95507-0

contextualists, rejecting the idea of moral action as rational action. My personal point of view can instead be characterized by the following assumptions:

1. Moral agency constitutes one kind of rational agency, i.e. a moral action cannot be irrational.
2. There are moral actions which are not 'rational' in the sense of the consequentialist theory of practical rationality.
3. An adequate theory of rational action is not consequentialist.

This point of view differs from the deontologist's critique insofar as it does not accept the idea that moral action is based on a specific form of rationality which has nothing to do with ordinary (instrumental or pragmatic) rationality. This deontologist separation of instrumental and moral rationality has its roots in Kant's distinction between pragmatic and moral imperatives.[2] Jürgen Haber-mas' communicative action vs. strategic rationality is a contemporary version of this Kantian dichotomy.[3] I am convinced that contrary to this dichotomy there is a unity of practical reason embracing moral and extra-moral reasons from a complex variety of different types of reasons partly based on principles, partly on obligations, partly on duties, partly on self-interest, partly on institutions etc., respectively, but I will not have the space here to delineate the constitutive types of good reasons for rational agency in general. Nevertheless, I hope that the argument is sufficient to show that consequentialism as a general theory of rational action fails and that this does not force us into deontologist dichotomies. The unity of practical reason can be saved without consequentialism.

Decision Theory and Consequentialism

Let us begin with the most forceful argument in favour of consequentialism. This argument says that if one accepts some quite minimal requirements of coherence constraining preferences, one is forced by deductive logical means to accept consequentialism. These requirements are well-known as the conditions of the utility-theorem. Let us call a preference-relation which meets these requirements 'Ramsey-coherent'.[4] Ramsey-coherent preferences are reflexive (as weak preferences: 'at least as good as'), they are complete, i.e. there is a (weak) preference between any pair of alternatives, and they are transitive, i.e. if x is preferred to y and y is preferred to z, then x is preferred to z. These three requirements are not sufficient to pass from coherence to consequentialism. The next step is to extend the assumed set of alternatives (or outcomes) X by including all probability distributions (or lotteries) over X Let us call this extended set X^*. Ramsey-coherence now requires additionally to the three just mentioned conditions that the preferences are reflexive, complete and transitive not only over X, but also over X^*, and that four additional requirements are met. The first is that an agent is indifferent between two probability distributions over X, if one can be transformed into the other according to the rules of the probability calculus. Secondly, if the rational agent is indifferent between a probability distributions* out of X^* and a specific alternative x out of X, then x^* and x can be substituted in any context without changing the preference

relation. Thirdly: if one element, let us call it 'x_b', is best and one alternative, call it 'x_w', is worst in X, then for each alternative x out of X there is a probability p such that the probability distribution $(px_b \ \& \ (1-p)x_w)$ and x are regarded as indifferent. The fourth requirement is that if there are two different probability distributions over two alternatives x and y out of X, then the probability distribution which assigns a higher probability to the better alternative, ranks above the other.

If a preference relation is Ramsey-coherent, then it can be represented by a real-valued function over X^*. This function is unique up to positive linear transformation. Representation here means that whenever an alternative x out of X^* is (weakly) preferred toy out of X^*, then the numerical value of x, which is assigned to x by this real-valued function u, is at least as big as the numerical value of y, which is assigned to y by u.

There has been some discussion whether these coherence-conditions are in fact adequate. For the sake of the argument we will assume that they are. If they are adequate, i.e. are indeed necessary conditions for the rationality of preferences, then the utility-theorem says that the rational agent is optimizing an utility-function. And if the rational person is optimizing an utility-function, then it may seem that this person is a consequentialist (or teleological) agent.[5] The possible consequences of the respective action are alternatives x out of X, which have an action-dependent probability greater than 0, these consequences have a specific (subjective) value $u(x)$, and the values of these possible consequences of the respective action are weighed by (subjective) probabilities. Beginning with some minimal requirements of preference-coherence, we get a full-blown consequentialist theory of rational action. If this move from coherence to consequentialism were well-founded, we had a strong argument in favour of consequentialism as a general theory of rational action. Still it is conceivable that after looking at these far-reaching implications of the coherence-assumptions we have a closer look and in the end dismiss at least one of them.[6] I do not exclude this possibility, but my argument is much more radical: This move from coherence-conditions to consequentialism is a *non-sequitur,* in other words: Coherentism about preferences does not justify consequentialism about actions.

Think of some ideal deontologist agent.[7] Such an agent might accept a variety of *prima facie* duties which in simple cases constrain the optimization of his personal interests. In more complex cases, i.e. in cases in which at least two of these *prima facie* duties are in conflict with each other, the agent is forced to apply some weighing procedure or to rely on his spontaneous intuitions. It depends on the content of the *prima facie* duties and – in complex cases – on the weighing procedure respectively the moral intuitions, whether this deontologist agent would act rational in a consequentialist sense or not. If we think of David Ross' list of *prima facie* duties, there can be no doubt that a person whose actions are guided by these *prima facie* duties is not acting rational in a consequentialist sense, because it is not the consequences (and their subjective value) alone which are relevant for choosing an action.[8] For example, if keeping one's promises is one of these *prima facie* duties, the past is intrinsically relevant for deciding. But the past is *per se* irrelevant for consequentialist rationality, since it is only the consequences of the respective actions what counts. If coherentism implied consequentialism, then the deontologist

agent's preferences had to violate at least one of the requirements of Ramsey-coherence. But I cannot see any reason why this should be the case. Why should the preferences of a person who is guided by the *prima facie* duty to keep one's promises be incoherent? We can certainly assume that David Ross' ideal moral person would have reflexive, complete and transitive preferences and that – if probabilites came into play – she would fulfill the four additional requirements named above. The same seems to be true for the ideal Kantian agent.

Most consequentialists might argue that moral deontology is incompatible with consequentialist rationality and therefore ill-founded. But this argument is convincing only if we assume more than modern utility theory does. If we already presuppose that the rational agent chooses his decisions exclusively on the basis of his subjective valuation of consequences (which are states of affairs causally – or probabilistically – connected with the respective action), weighing the consequences by the measure of his subjective probabilities, then the incompatibility with the deontologist agent is obvious. Therefore, if consequentialist rationality were implied by Ramsey-coherence, then we had to assume that the preferences of the ideal deontologist agent would violate at least one of the coherence-conditions. But since consequentialist rationality is not implied by Ramsey-coherence, the incompatibility of consequentialist rationality and deontology does not imply that the deontologist agent is Ramsey-incoherent.

The incompatibility of consequentialist rationality and acting on deontological reasons can be made more explicit if we take a closer look at the notion of a 'consequence'. Loosely speaking, every property of an action can be interpreted as a possible consequence. If an action fulfills a promise, say, this fact can be understood as one of its consequences. Such a wide notion of 'consequence' would make the idea of consequentialist rationality trivial or empty. The intuition on which the theory of consequentialist rationality is based is that we look at the causal effects which an action has on the history of the world and that we value an action as right or wrong exclusively with regard to these causal effects. This intuition can only be saved if the notion of consequence applied in this context implies that all relevant aspects of consequences can be described using unhistoric predicates. Classical utilitarian consequentialism is a good example. The evaluation of the consequences is exclusively determined by the sum of individual well-beings, whereas 'well-being' is to be interpreted in such a way that it characterizes a subjective state of mind. Historical information might be necessary to form rational expectations over consequences, but consequences themselves can be described without using historic predicates. History has a causal influence on consequences, but the value of a given consequence is independent from history. This explains in more abstract terms why deontological agency in many cases cannot be consequentialist. To help a miserable person to feel better is a good reason for action, and there might exist a duty to help the miserable. If we have only this duty in mind, the only relevant information to judge the adequacy of an action would concern consequences, more precisely the consequences regarding the subjective state of the miserable person. But we have good reasons to follow other duties such as being truthful, keeping promises etc., which require more information than those concerning the consequences of the respective action only.

Consequentialism and the Moral Point of View

In the last section I have shown that we have to discriminate sharply between a coherentist theory of rational preference and a consequentialist theory of rational action. For the latter, there is a primacy of the (subjective) valuation of consequences which motivates the (consequentialist) person to choose an action out of her options if and only if this action optimizes the consequential value regarding her subjective probabilities. If moral agency is to be rational, morality can only be constituted by a specification of how consequences are to be evaluated. It is a constitutive trait of consequentialism that if the subjective valuation of consequences is given, then the rational action is well-determined except for cases of indifference. Aspects of universalizability e. g. are irrelevant for consequentialist rationality.

The Kantian theory of practical reason says that whatever your interests are and however you value consequences on the basis of these interests, you are acting irrational if following these interests is not universalizable. The Kantian theory of practical reason includes two steps: first you form the subjective rules (maxims) on which you want to act, second you check the interpersonal compatibility, i.e. universalizability of these maxims. The second step imposes constraints on following your personal interests. Consequentialism on the other hand is a one-step theory of practical rationality, and this is one of the main strengths of this account. You have to know how the person values consequences, and then you can tell the person what is rational for her to do (dependent on her subjective probabilities). Therefore, within a consequentialist conceptional framework the whole burden of moral evaluation rests on the individual value-function over consequences.

Most practical philosophers agree that the moral point of view requires some form of impartiality, but there are many different opinions about the content and the range of morally grounded impartiality. Deontologists accept the variety of different subjective points of view if these are action-guiding only within some impartial constraints. ‘Impartiality by universally acceptable constraints’ might be taken as the basic formula of the deontologist theory of practical reason. For the consequentialist the impartiality of the moral point of view must become part of the value-function over consequences which motivates the rational person’s actions. The most radical consequentialist inclusion of the moral point of view requires each rational and moral person to adopt the same subjective value-function over consequences. The right action is then defined as optimizing this interpersonal invariant value-function over consequences. The question of how this interpersonally invariant value-function is to be constituted has different answers, ranging from classical utilitarian theory, which identifies this value-function with the total sum of happiness, to modern theories, identifying this value-function with an aggregate of individual ‘personal’ preferences, giving each individual preference-relation equal weight.[9]

Let us call a theory of this kind requiring an interpersonally invariant value-function as the only adequate expression of the moral point of view ‘strict ethical consequentialism’. The theoretical design of strict ethical consequentialism is simple and elegant; it adheres to consequentialism as the general theory of rational

action on the one hand and it incorporates the moral point of view by means of withdrawing interpersonal differences regarding the action-guiding evaluation of consequences. The problem with strict ethical consequentialism is that it obviously requires too much. Not only that it is supererogatory because nobody can be expected to choose his actions such that everybody's interests in the world are equally fostered, but – more fundamentally – the idea of a society in which every individual maximizes the same value-function disregards the necessary and valuable complexity based on individual differences. A good society is constituted by individuals with different projects and bindings, different life-plans and virtues, different values, hopes and fears.

If strict ethical consequentialism requires too much and the ordinary *homo oeconomicus* model of optimizing personal interests is incompatible with the moral point of view, then it seems reasonable to choose a 'middle road'. The middle road adheres to the consequentialist theory of practical rationality and to the idea that moral agency is rational agency, but it dismisses the inter-personally equilibrizing tendency of strict ethical consequentialism. Samuel Scheffler's agent-centered prerogative is an example. It is still a consequentialist theory insofar as an action is right if and only if it optimizes the value of (expected) consequences, but it allows the agent to give unequal weight to the interests of different persons. It is allowed (not required) that the agent gives more weight to his personal interests than to the interests of others. The moral point of view has adopted a more humane stance. We can think of a continuum beginning with the *homo oeconomicus* model of self-oriented maximization and ending with strict ethical consequentialism. Far from being a 'rejection of consequentialism', this proposal – even if it might imply many problems in detail – seems to save consequentialism as a general theory of rational action because it incorporates the moral point of view without destroying individual differences.

To summarize: consequentialism as a general theory of rational action is incomplete if it is not able to account for good moral reasons for action. One possible account is the one of strict ethical consequentialism. Since strict ethical consequentialism is incompatible with interpersonal differences, it is inadequate for anthropological reasons. But there seems to be a vast space between the self-oriented *homo oeconomicus* on the one hand and the strict ethical consequentialist on the other to find an adequate reconciliation of consequentialism and the moral point of view.

Consequentialism and Differences

The consequentialist *homo oeconomicus* accepts every kind of interpersonal differences. For the *homo oeconomicus* these differences are differences of interests, however, they might be laden with moral values. On the theoretical level this account excludes moral reasons for action. Strict ethical consequentialism on the other hand includes good moral reasons for action in the most demanding form of prescribing one interpersonally invariant value-function (over consequences) for each individual. Weaker forms of (ethical) consequentialism try to include good moral reasons

for action in a weaker form, such as allowing every person to give special weight to her own interests, but limiting this 'agent-centered prerogative' in some way or another. Think for example of a weighing-factor which allows to aggregate personal interests only within a range in which no one's interest counts more than five times as much as the interest of any other person.[10] Proposals of this kind result in different personal value-functions over consequences. They allow for different personal projects and plans of life in general. But when consequentialism tries to include good moral reasons in a general account of rational action without going all the way to strict ethical consequentialism, the failure becomes apparent. If a normative theory allows for interpersonally different valuations of consequences, it has to say how to solve the problems of individual rights, collective decisions and cooperation. None of these three problems can be adequately answered within the consequentialist framework. Some well-known decision-theoretic results can help to make this argument transparent and short.

1. *Individual Rights*

Consequentialism is bound to collective choices which are pareto-inclusive. Some social state is better than another if at least one person is better off in the first and no person is worse off in the latter. Even the traditional *homo oeconomicus* version of consequentialism is closely linked to the theory of the ideal market in which individual optimizers realize a social state which is pareto-efficient.[11] Versions of consequentialism which are not bound to one single type of good reasons for action which are based on self-interest, i.e. versions of consequentialism which try to include genuine moral reasons for individual action, *a fortiori* cannot accept pareto-inefficient social states as the result of ideal rational behaviour. Whatever hybrid versions of consequentialism offer as a compromise between the personal point of view and impartiality, the resulting normative collective choice rule has to be pareto-inclusive. But whenever individual differences are allowed for, it seems reasonable to attribute individual rights, too. Individual rights protect the personal sphere, they are part of those conditions which are needed to live a personal life. But as Amartya Sen has proven in (1970), it is not possible to have both conditions fulfilled by a reasonable collective choice rule. The deontologists' insistence on constraints for optimization seems to make sense again. It is not possible to combine consequentialism, interpersonal differences, individual rights and efficiency.[12]

2. *Collective Action*

Collective action requires some commonly accepted procedure of aggregation. Ideal rational consequentialists optimize the expected value of consequences in every single decision. To participate in a collective action and the form how to participate are decisions like any other. A well-known theorem of collective choice, independently proven by Allan Gibbard and Mark Satterthwaite[13], has far-reaching implications for hybrid versions of consequentialism, since we cannot assume that some kind of simple interdependency of preferences, as it is proposed by hybrid consequentialism, would suffice to make a reasonable collective choice rule strategy-proof. But if reasonable collective choice rules are not strategy-proof in a

collective of hybrid consequentialists, then we face a basic instability in a society of ideal moral and rational consequentialists. Collective action requires agents who under certain circumstances refrain from optimization and stick to the rules of the game.

3. *Cooperation*

Genuine cooperation is a type of action for which the following holds: if each of the (two or more) participants optimized their respective value-function over consequences, each participant would be worse off than if all participants would choose the cooperative action.[14] If we characterize cooperation in this way, genuine cooperative action is bound to situations of the prisoners' dilemma type. But if we use the usual characterization of the prisoners' dilemma, cooperative action is rendered irrational, i.e. Ramsey-incoherent. If a two-person prisoners' dilemma is defined by the following preferences: *A* (row-chooser): *DC* > *CC* > *DD* and *B* (column-chooser): *CD* > *CC* > *DD* > *DC,* then to decide for action *C* is obviously incompatible with these preference-relations for *A* and *B* respectively if their decisions are independent from each other. Strict ethical consequentialism excludes such combinations of preference relations. It excludes the possibility of a prisoners' dilemma situation because the action-guiding preferences of ideal strict ethical consequentialists are interpersonally invariant. But the whole rest of the consequentialist spectrum allows for interpersonal differences and cannot exclude PD-situations. If the consequentialist theory of rational action includes moral reasons – other than the standard *homo oeconomicus* model –, it cannot at the same time exclude one of the most central moral reasons, and this is cooperation.

How is it possible to cooperate without having incoherent preferences? The answer is that we have to enrich the description of the situation in which cooperative action is possible. Let us even grant for a moment that the action itself has no intrinsic value whatsoever and that the only relevant properties of the feasible actions concern the consequences of these actions. The only additional assumption which we need for our argument is that the preference structure of the situation is part of the information on which the individual reasons for action rest. Within the orthodox consequentialist framework we would be forced to describe an agent willing to act cooperatively under the (necessary) condition that she can expect the other agent to be willing to cooperate conditionally, too, by the following preference structure: *CC* > *DC* > *DD* > *CD,* which constitutes the assurance game in case the agent *B* has analogous preferences. These (assurance game-) preferences are compatible with the assumption that the individual preferences over consequences have not changed at all.[15] In other words: the PD-structure can survive if we look at the participants' preferences over consequences whereas at the same time they reveal a cooperative attitude by assurance game preferences regarding the possible strategy combinations. This cooperative attitude might be motivated exclusively by two bits of information: the own and the other person's evaluation of consequences (which constitute a PD) and the knowledge of the others' conditional cooperative attitude. If we go one step further and assume an agent who cooperates unconditionally (other-regarding preferences), we can even drop the second bit of information. Obviously

we have to drive a wedge between choice and preference in order to reconstruct cooperative action. The cooperative agent does not optimize (individually) her evaluation of consequences.

Weighing Reasons

In many everyday situations we have good reasons to respect individual rights, stick to some accepted rules of collective action or to cooperate. In these cases we refrain from optimizing consequences. But still we try to be intra- and interpersonally coherent. Ramsey-coherence is a minimal requirement of being coherent, i.e. living a reasonable life. Living a reasonable life requires weighing reasons. Different types of reasons refer to different properties of actions and their contexts. In some rare cases reasons exclusively refer to consequences. In other cases they refer to obligations which have arisen due to past actions, duties which are bound to social roles, individual rights which protect autonomy and universal principles as constraints. To include these reasons adequately, it is necessary to enrich the conceptual frame of the theory of rational choice beyond consequentialism. Ramsey-coherence is the result of weighing reasons, not weighing goods. Ramsey-coherence is a core element of the unity of reason.

Appendix

The purpose of this appendix is to state more formally some of the technical concepts and results referred to and used in the arguments of the main body of the paper.

1. *Ramsey-coherence*

Let $X=\{x_1,...,x_2\}$ be a set of outcomes and let $R \subset X \times X$ be a preference relation of some individual over X (where '$(x, y) \in$ R' is to be read as 'the individual weakly prefers x over y'). R is called an *ordering* iff the following three conditions are satisfied:

Reflexivity. $\forall x \in X$: $(x, x) \in R$. – Every outcome is at least as good as itself.
Completeness. $\forall x, y \in X$: $(x, y) \in R \vee (y, x) \in R$. – Every pair of outcomes can be ordered according to the weak preference relation.
Transitivity. $\forall x, y, z \in X$: $((x, y) \in R \wedge (y, z) \in R \rightarrow (x, z) \in R)$. – If an outcome x is weakly preferred over y and y is weakly preferred over z, then x is weakly preferred over z.

Let $p_1,..., p_n \in [0,1]$ be nonnegative real numbers smaller or equal than 1 with $P_1 + ... + p_n = 1$. Then the n-tupel $(P_1/x_1 \,\&\, ... \,\&\, p_n/x_n)$ is called *a prospect* (hence, a prospect is a probability distribution over outcomes). X^* denotes the set of all prospects and we assume the preference relation R to be extended over X^*. R is called *Ramsey-coherent* iff it is an ordering and, additionally, the following four conditions are met:

Continuity. If for the individual the best element from X is x_b and the worst element is x_w (and the individual strongly prefers x_b over x_w – i.e. is not indifferent between

all elements from X), then for each x in X there is a probability $p \in [0,1]$ such that the individual is indifferent between x and the prospect $(p/x_b \ \& \ (1-p)/x_w)$.

Monotonicity. Let $A = (p/x \ \& \ (1-p)/y)$ and $A' = (p'/x \ \& \ (1-p')/y)$ be two prospects. If the individual strongly prefers x over y, then $(A, A') \in R \leftrightarrow P \geq P'$.

Reduction. Prospects can be manipulated according to the probability calculus without changing the individual's preferences between them.

Substitution. If the individual is indifferent between an outcome x and a prospect A, then x is substitutable by A in every prospect A' in which x occurs withour changing the individual's preferences.

It can be shown that if the individual preference relation over X^* is an ordering and additionally meets these four constraints, then there is a linear function u (which is – modulo positive linear transformation – uniquely determined) from the set of prospects into the real numbers which cardinally represents the individual's preferences: $\forall x, y \in X^*$: $u(x) \geq u(y) \leftrightarrow x \geq y$

2. *The Liberal Paradox*

Let again $X=\{x_1,...,x_n\}$ be a set of alternatives and let a set $K=\{1,2,...\}$ be a group of individuals (the 'collective'). *A preference structure* is a function g mapping individuals into preference relations: $g\colon K \rightarrow \mathrm{Pot}(X\times X)$ (where $\mathrm{Pot}(X\times X)$ denotes the power set of X). For each individual $i \in K$, $g(i)$ is the weak preference relation of i over X. Let the set of all preference structures be denoted by G. The strong individual preference relation for individual i is denoted by $g(i)$.

A choice function f mapping preference structures (in which all the individual preference relations are orderings) into collective orderings is called a *social decision function.* We define two properties a social decision function may or may not have.

Strong Pareto Principle. $\forall g \in G \quad \forall x, y \in X: \ [\forall i \in K : (x, y) \in g(i) \wedge \ \exists j \in K : (x, y) \in g(j) \rightarrow (x, y) \in f(g)]$. – If every individual regards an alternative x to be at least as good as an alternative y and at least one individual stricdy prefers x over y, then x should be collectively strictly preferred over y.

Liberalism. $\forall i \in K \ \exists x, y \in X \ (x \neq y) \ \forall g \in G : [((x, y) \in g(i) \rightarrow (x, y) \in f(g)) \wedge ((y, x) \in g(\mathrm{i}) \rightarrow (y, x) \in f(g))$. – For every individual there are at least two alternatives over which the individual is decisive (when 'decisive' means that the preferences of the respective individual are to become the social preferences).

Sen's theorem states that there is no social decision function simultaneously satisfying the Liberalism condition and the Strong Pareto Principle.

3. *The Theorem of Gibbard and Satterthwaite*

Let X be a set of alternatives, K the collective, G the set of all possible preference structures, G^s the set of preference structures consisting only of strict individual preferences, and g a preference structure.

If $g = < g(1),...,g(n) >$ and $g' = <g'(1),...,g'(n') >$, then we define $g|g'(i) := < g(1),...,g(i-1), g'(i+1),\ldots,g(n) >$. For any individual i let $g^s(i)$ be the set of possible

(strict) individual preference relations for *i*. A function f: $G \rightarrow \text{Pot}(X \times X)$, $g \mapsto R = f(g)$ mapping preference structures into preference relations over X is called a *rule of aggregation.* A *choice function* is a function $a_{f(g)}$ mapping subsets of X into choice sets according to the collective preference ordering given by an aggregation rule f applied to a preference structure g. A choice function is said to fulfill the condition of *strategy-proofness* iff

$$\neg\exists g \in G^S : \exists i \in K : \left[\exists g'(i) \in g^S(i) \wedge \left(a_{f(g|g'(i))}(X'), a_{f(g)}(X') \right) \in g(i) \right].$$

If a choice function is strategy-proof, then it is not optimizing for an individual in terms of his preferences $g(i)$ to feed other preferences into the collective aggregation process than his 'real' ones.

The theorem proven by Gibbard and Satterthwaite states that there is no strategy-proof choice function (except for the dictatorial rule). Every non-dictatorial rule includes an incentive for consequentialist optimizers to manipulate (and counter-manipulate). This results in a basic instability.

4. *The Prisoners' Dilemma*

A prisoners' dilemma (of the standard type) is an interaction situation involving two agents (labelled *A* and *B*) each facing two possible alternatives, *C* and *D* ('*C*' for 'cooperation' and '*D*' for 'defection'):

		B	
		C	*D*
A	*C*	*CC*	*CD*
	D	*DC*	*DD*

The *prisoners' dilemma* (PD) is defined by the preference structure

A: *DC* > *CC* > *DD* > *CD*
B: *CD* > *CC* > *DD* > *DC*.

The *assurance game* (AG) is defined by the preference structure

A: *CC* > *DC* > *DD* > *CD*
B: *CC* > *CD* > *DD* > *DC*.

The *other-regarding game* (OR) is defined by the preference structure

A: *CC* > *DC* > *CD* > *DD*
B: *CC* > *CD* > *DC* > *DD*.

References

Broome, J. 1991. *Weighing Goods.* Oxford: Basil Blackwell.

Gauthier, D. 1986. *Morals by Agreement.* Oxford: Oxford University Press.

Gibbard, A. 1973. Manipulating Voting Schemes: A General Result. *Econometrica* 41: 587–601.

Habermas, J. 1981. *Theorie des kommunikativen Handelns.* Frankfurt a.M.: Suhrkamp.

Harsanyi, J.C. 1955. Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. *Journal of Political Economy* 63: 309–321.

Kant, I. 1785. *Grundlagen zurMetaphysik derSitten.* Transi, by H.G. Paton, *Groundwork of the Metaphysics of Morals.* New York 1956.

Kern, L., and J. Nida-Rümelin. 1994. *Logik kollektiver Entscheidungen.* München/ Wien: Oldenbourg.

Margolis, H. 1982. *Selfishness, Altruism, and Rationality.* Cambridge: Cambridge University Press.

McClennen, E.F. 1990. *Rationality and Dynamic Choice.* Cambridge: Cambridge University Press.

Nida-Rümelin, J. 1991. Practical Reason or Metapreferences? An Undogmatic Defense of Kantian Morality. *Theory and Decision* 30: 133–162.

Nida-Rümelin, J. 1993. *Kritik des Konsequentialismus,* paperback ed. 1995. München: Oldenbourg.

Ross, D. 1930. *The Right and the Good.* Oxford: Oxford University Press.

Satterthwaite, M.A. 1975. Strategy-Proofhess and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions. *Journal of Economic Theory* 10: 187–217.

Scheffler, S. 1982. *The Rejection of Consequentialism.* Oxford: Clarendon Press.

Scheffler, S. (ed.). 1988. *Consequentialism and Its Critics.* Oxford: Oxford University Press.

Sen, A.K. 1970. The Impossibility of a Paretian Liberal. *Journal of Political Economy* 78: 152–157.

Sen, A.K. 1977. Choice, Orderings, and Morality. In *Practical Reason*, ed. S. Körner, 54–67. Oxford: Oxford University Press.

Vallentyne, P. 1987. The Teleological/Deontological Distinction. *The Journal of Value Inquiry* 21: 21–32.

Vallentyne, P. 1988. Teleology, Consequentialism, and the Past. *The Journal of Value Inquiry* 22: 89–101.

Why Rational Deontological Action Optimizes Subjective Value

Abstract In present day philosophy there are two competing views regarding practical rationality: (1) Decision and game theory and economic theory have developed a theory of rational decision which has proven to be fruitful in many areas of social science. Practical philosophy should work with thatparadigm (2) Economic theory and decision theory do not have an adequate account of practical rationality. The homo oeconomicus model is – at best – one perspective which competes inter alia with philosophical accounts of practical reason.

In this article I try to show that these two seemingly opposing views arein fact compatible. I argue that consequentialism is an inadequate account of rationality because rational action is deontological in character. Nevertheless the decision theoretic conceptual frame should not be given up. Deontology and decision theory can be made compatible via comprehensive description of action. The conceptual frame of decision theory should be interpreted as coherentist, not consequentialist. With this intertretation deontological action,if rational, maximizes subjective value.

Introduction

Does rational deontological action optimize subjective value? Can one be a deontologist and at the same time adhere to decision theory as an all-embracing theory of practical rationality? I think the answer to both of these questions is *yes*. The reasons for it are given in this article.

There are two basic intuitions which frame the bigger part of practical philosophy and which seem to be incompatible. One intuition is *teleological* or more specifically *consequentialist* according to which rational action optimizes its consequences. The other intuition is *deontologicalox rule-oriented* according to which rational action is guided by certain rules. I am a deontologist, **I** think that consequentialism is an inadequate theory of ethics and rationality alike, but at the same time **I** am convinced that rational action maximizes subjective value.

The reader probably thinks that the following two assumptions cannot be true simultaneously

(1) Consequentialism as a theory of *rationality* is false
(2) Rational action maximizes subjective value respectively
(1′) Consequentialism as a theory of *morality* is false
(2′) Moral action maximizes subjective value.

Published in *ProtoSociology – An International Journal of Interdisciplinary Research*, Vol 21, 2005, pp. 182–193.

I try to show in this article that this alleged incompatibility does not exist. The teleological intuition can be upheld not in the original *consequentialist*, but in a weaker *coherentist* form. A deontological (anti-consequentialist) theory of rationality and morality, as I think it to be adequate, is compatible with coherent preferences and beliefs and fits therefore into the decision theoretic conceptual frame, if interpreted in a coherentist manner.

If one takes moral actions to be rational actions, there is a close link between ethics and the theory of practical rationality.[16] The best-developed theory of rationality, however, is decision theory (including game theory), and decision theory is a consequentialist theory of rationality, or so it seems. Most ethical deontologists share the common philosophical belief that decision theory is about prudence (or strategic action), whereas morality establishes constraints on prudential optimizing. Deontologists, therefore, tend to be hostile towards the idea of taking decision theory to be the core of a general theory of practical rationality and towards the attempt to integrate moral actions into its conceptual framework.

If the decision theoretic framework cannot be dismissed in order to define rational action, it seems that ethical consequentialism has strong arguments in its favour. Ethical axiology would be about moral values and ethical rationality about how to optimize moral values. Utilitarianism and decision theory which both are closely interlinked in the history of ideas would then still be dependent on each other from a systematic perspective, too. I will instead argue that rational deontological action is compatible with standard decision theoretic axioms. *Rational deontological action optimizes subjective value.*

The Wedge Between Choice and Preference

Decision theory works with the assumption that the rational person reveals two basic propositional attitudes in acting: *subjective preferences* and *subjective probabilities.* It attributes two real-valued functions to the rational actor: the function of *subjective probability* and the function of *subjective value.* This attribution is based on the comparative concepts *preferring* and *expecting* – person *i prefers p to q* and *i takes p to be more probable than q.* It is not possible to attribute one of these two functions independently from the other, and insofar the attribution of the two functions is interlinked.

Is decision theoretic rationality consequentialist? The standard applications of decision theory are indeed consequentialist insofar as *subjective value* is defined as *subjective value of consequences,* and insofar *rationality* is defined as *maximizing expected subjective value of consequences,* or in short as: *optimizing consequences.* Conceptually this consequentialism is made explicit in Savage's model, for example, but not in Jeffrey's holistic account.[17] If we think that moral actions should be rational actions, morality should be defined as maximizing expected subjective value of a *specific kind.* It would be the task of *ethical theory* to develop criteria under which a subjective value function seems morally acceptable.

What, then, if you are a deontologist, that is, if you think that deontological criteria cannot be dismissed in ethical theory? Under the assumptions that the decision theoretic framework can not be dismissed either and that decision theory commits one to a consequentialist view on rational agency, the application of deontological criteria seems to suggest the following options: either the idea that moral action is rational has to be abandoned, or two different kinds of rationality have to be adopted: one for moral and one for extra-moral action – in the latter case, decision theory would have to be taken to deal with extra-moral actions only.

The first option appears unattractive: "You should do x, but it is irrational to do x!" is not a convincing moral imperative. The second option would imply that decision theory is not a general theory of rational action. In this article, I shall not be concerned with the first option (we dismiss it as implausible without further argument), but focus on the second: Are we to accept that there are two quite different kinds of rationality? Is it true that deontological action does not fit into the decision theoretic frame?

Decision theory does not begin with postulating probability and value functions. Rather decision theory begins with describing *basic properties* of preference relations. Some of these properties can be interpreted as analytically tied to the notion of preference, others might be understood as synthetic (and normative insofar as they, taken together, constitute rationality or are at least necessary conditions of rationality.). Here – as elsewhere – there seems to be no clear-cut borderline between *analytic* and *synthetic.* On the one hand, it seems that *reflexivity* of preferences is an analytic property, whereas *transitivity* of preferences rather appears to be a synthetic (and normative) requirement. On the other hand: if a person has preferences regarding three well-determined, well-described and well-known alternatives which are intransitive, one may well ask the person, whether she *really* has those preferences. If we think of preferences as being revealed in action and in verbal expression, violations of transitivity or other properties of rational preference relations may result in uncertainty concerning the reliability of the attribution of these preferences in the first place.

This phenomenon indicates that there is no precise demarcation line between analytic and synthetic properties of rational preference relations.

Is there any reason to assume that deontological action violates one of these properties of rational preferences (whatever their status may be), e.g. transitivity? This formulation is an elliptic version of: is there any reason to assume that actions which are performed for deontological reasons reveal preferences which violate the transitivity condition? It makes things easier if we circumvent the notion of *action* by asking: is there any reason to assume that *preferences* which are based on deontological rules violate the transitivity condition?

Let us illustrate the point by referring to the deontological rule THOU SHALT NOT STEAL. The conflict between deontology and consequentialism is apparent: Even if stealing in some cases might have overall good consequences, you are not allowed to steal. This remains true also if we concede that in cases of extremely bad consequences e.g. starving, stealing might be justified.

Now think of the following case. There are three possible states of the world *x,y,z.* We assume that these three states of the world are certain[18] consequences of three alternative feasible actions *u,v,w* respectively. The person in question prefers *x* to *y* and y to z. The subjective value function represents these preferences adequately. Since probability is not involved, maximization of expected subjective value would result in choosing action *u.* But let us assume that *u* is an act of stealing, whereas *v* and *w* are no acts of stealing (and are not forbidden by other deontological rules). In this case it seems that *v* should be the preferred action. The situation is such that it is not possible to optimize consequences and at the same time conform with the deontological rule in question.

We can assume that the rational person has *transitive* (and in general *coherent)* preferences regarding states of the world as well as regarding actions. Both kind of preferences should be coherent, if the person is assumed to be rational. But the rational person's preferences regarding states of the world do not necessarily determine the person's preferences regarding actions. This is not implied by the axioms of the utility theorem. It is implied only by specific interpretations, that go beyond what is requiered by the axioms themselves.

Reflexivity of weak preferences is given trivially (in our examplee there are no weak preferences involved) and *Completeness* of preferences is a reasonable postulate for a deontologist and a consequentialist alike. Since the rational deontologist (like the consequentialist) cannot refuse to act, she reveales her preferences regarding any feasible alternative. There is no reason to assume that a rational person with deontological motivations gets sometimes caught in holes of inactivity. The *Completeness Axiom* is valid for preferences motivated by deontological reasons, too.

One widespread form of presenting decision theory identifies actions with prospects, in the sense of probability distributions over the sets of their consequences and takes consequences as states of the world (or sets of states of the world). Therefore it identifies *u* with x, *v* with *y* and *w* with *z.* Within this conceptual framework there, indeed, is no logical space for deontological rationality. It is excluded by conceptual means that a rational person has preferences as we assumed above. *This conceptual exclusion of deontological rationality is unacceptable for any adequate theory of practical reasons and it is unnecessary in order to preserve the conceptual frame of decision theory.*

It makes perfect sense that a person, when asked which state of the world she prefers, answers that she prefers *u* to *v* and *v* to *w* while she respects the Seventh Commandment and therefore prefers action *y* to action *x.* There is no reason to think that everybody, who respects this commandment is irrational. A rational person can prefer the state of the world *x* to *y* and *y* to *z* and at the same time prefer the action *v* to the action *u*, because *u* would violate the Seventh Commandment and *v* wouldn't.

With this example in mind, the flaw in standard consequentialist decision theory is apparent: Preferences regarding states of the world are insufficient to determine rational action. *We have to drive a wedge between choice (of actions) and preference (between states of the world).* The adequate reaction to this flaw is to avoid any conceptual dependency on consequentialism. Even the friends of consequentialism

cannot wish that consequentialism is true by definition. If consequentialism is true it is true by synthetic normative argument. If it is wrong, it is wrong by synthetic normative argument. If the *open question argument* is of any value at all, it can well be applied to this case: "Action *x* has *u* as its consequence and *u* is the best of the three possible consequences *u,v,w,* but is *x* rational for somebody who accepts the Fifth Commandment?" is a reasonable question, that cannot be ruled out by conceptual means.

We are therefore forced to draw a conclusion that has far reaching implications for any adequate theory of practical rationality: *Not all actions can be represented by their prospects* in the sense of probability distributions over the set of their consequences, if "consequences" are understood as "states of the world" or "sets of states of the world". Actions change probabilities of propositions (understood as sets of states of the world), this is how actions intervene in the course of history. The prospects of an action can be quite relevant to determine whether it is rational. It is, however, an essential trait of consequentialist persons only that for them actions can be represented by their prospects (understood as described above). But the preferences of a rational person regarding prospects (sets of state of the world) *can* differ from the same persons preferences regarding the actions that cause these prospects. *It is not universally true that one can substitute actions for prospects within a rational persons preferences.* It is not true in case of *deontological* rational actions. *Deontological actions can not be represented by their prospects in a rational person's preference relation.*The rational and deontologically motivated person drives a wedge between choice (of actions) and preference (of states).

Deontological actions are motivated by deontological reasons. An action complies with deontological reasons *under a description that makes this compliance apparent.* But this is exactly why deontological actions cannot be represented by their prospects. To describe an action by its prospects alone excludes that under this description it is rationally motivated by deontological reasons. An action can comply with deontological reasons only under a *comprehensive* description as opposed to a *reduced* description by sole reference to its prospect.

Ramsey-Coherence

The question remains whether it makes any sense to apply the decision-theoretic framework to *comprehensively described rational deontological actions.* In order to answer this question we have to check whether the basic axioms of decision-theoretic rationality are applicable to preferences regarding comprehensively described rational deontological actions. If *u,v* and *w* are three comprehensively described rational deontological actions, a rational person who prefers *u* to *v* and *v* to *w* prefers *u* to *w.* This requirement of *Transitivity* should be met not only in case the rational persons preferences are based on consequentialist, but also if they are based on deontological reasons. Why should deontological reasons go against transitivity? Why should a rational agent develop intransivitive preferences only because

he is motivated by deontological reasons? Why should deontological motivations lead to intransitivity?

I think that there are no reasonable answers to these questions. Moreover, if a set of deontological reasons justified intransitive preferences this set of deontological reasons *ipso facto* would be shown to be incoherent: Transitivity of preferences is a requirement of practical coherence that can not be dismissed if we move from consequentialist reasons to deontological ones. Likewise we can assume that the decision-theoretic requirements of *Reflexivity* and *Completeness* are met by the rational person's preferences within this set of comprehensively described rational deontological actions.

Problems seem to arise if we include probabilistic considerations. We are used to interpret prospects that are assigned to actions in a consequentialist manner. In doing *a* I expect the consequence *c* with probability 0.2 and the consequence *c*' with probability 0.8. But what means "consequence" here? The verbal expressions we use for describing consequences are underdetermined regarding this question, because the answer is dependent on what we take to be relevant for a rational preferential ranking of actions. It might be understood as a 'consequence' of doing *a* that I keep my promise or that I obey the Seventh Commandment, such a 'consequence' though is not a state of the world or a set of states of the world, in a consequentialist sense. It is a proposition, but a proposition that can be true or false without changing the state of the world. It might in this sense not have any 'consequences'. If you take 'consequence' as being – more specifically – characterized by the state of the person's mind[19], this remains true a fortiori. The decision-theoretic axioms, however, should not determine what is relevant for establishing a rational preference-relation.They merely should constrain the form of it in order to secure practical coherence, they require *inter alia* that the preference-relation is an ordering.

As explained in the last section deontologically motivated preferences require *comprehensively described* actions. The reason is that the prospects of actions in the usual consequentialist sense do not include enough information for the deontological agent to rank his actions. But if this is true it seems that we loose any ground for including probabilistic considerations. And if probabilistic considerations were excluded we could not apply decision-theoretic axioms that take probabilistic considerations into account (the monotonicity and continuity axiom). In this case the decision-theoretic framework would collapse. It would become impossible to assign a well-determined (up to positive linear transformation) function of subjective value to the rational deontologist. The transformation of preference into value would not succeed any more.

Luckily there is a way out. In order to find it let us remember what Elisabeth Anscombe[20] and many philosophers of action after her described.Actions result in other actions and it is often not clear to the actor in which actions his actions result. If I press the switch it is not sure whether this action results in the action to switch on the light, because there might be a defect in the electrical wire system. If I switch on the light by moving the switch, it may be that I frighten my wife by switching on the light or that I prevent the burglar to commit a crime. In many cases, it is enough

to know which effects my actions have on the state of the world, but in deontologically relevant situations this is not enough. Some action of mine *a* may with non-zero probability result in an action *b* which means that I break my promise. If I am a deontologist I may think that breaking promises is not right, even if it has good consequences regarding the state of the world. Many persons, myself included, have such a normative belief, and we should not exclude that some of these persons are nevertheless rational. But then it is relevant for such a person to know the probability of her breaking the promise as a result of doing *b* by doing *a.*

If I accomplish an action *a* and this action will result in action *b* or action *c* and if *a* is of no intrinsic value, I should be indifferent between *a* and the respective probability distribution between *b* and *c* which *a* brings about. The *a*-conditioned probabilities of *b* and *c* should determine my decision. This remains true also if *b* and *c* are comprehensively described and valued under deontological criteria. The probabilistic axioms of decision theory describe how the *a*-conditioned probabilities of *b* and *c* should determine my decision. A rational person has coherent preferences regarding probabilistic relations between actions.

If I prefer action *u* to *v* and by doing *a* I accomplish either *u* or *v* and by doing *b* I also accomplish either *u* or v and if *a* and *b* are deontologically neutral in the sense explained in the last paragraph, and if *u* is more probable if I do *a* than if I do *b,* then it would be irrational to do *b.* To do *a* in such cases is a minimal requirement of coherence that deontologically motivated rational preferences should fulfil. Let us call this assumption the *Extended Monotonicity Axiom.*

If I prefer action *u* to *v* and *v* to *w* for deontological reasons, then there is a unique probability distribution on *u* and *w* such that I am indifferent between this prospect and *v.* This prospect can well be caused by an action although the consequences are again actions. This is another minimal requirement of coherence that deontologically motivated rational preferences should satisfy. Let us call this assumption the *Extended Continuity Axiom.*

If we add the assumption that the preferences of a deontologically motivated and rational person are *complete (Extended Completeness Axiom)* and *invariant* according to the *Probability Calculus (Extended Substitution Axiom),* we can conclude by simple mathematical deduction that her preferences can be represented by a subjective value function.[21] *A rational person following deontological reasons optimizes subjective value,*[22] because her preferences adequately described are Ramsey-coherent.[23]

Remarks on Interpretation

This result needs some interpretation with which I want to end this article.

1. It goes without saying that our result does not show that consequentialism is indispensable. John Broome, however, used an argument which has some similarity to mine in order to show that consequentialism is conceptually unavoidable.[24] If the reasoning above is correct, this is a *non-sequitur.* The mistake

is due to a confusion of *optimizing subjective value* and *optimizing consequences.* If consequences are states of the world (or sets of states of the world, or propositions referring to states of the world) then optimizing subjective value is compatible with *not* optimizing consequences. Following deontological rules is compatible with optimizing subjective value, even more: following deontological rules *rationally* requires optimizing subjective value.

2. The rational deontologist has coherent preferences for actions and states of the world alike. The deontologist differs from the consequentialist in that her preferences in general cannot be *reduced* to her preferences between states of the world. The rational deontologist is not hostile to consequentialist reasons, in many cases there are no deontological reasons which have to be considered in order to take a good decision. These cases allow to extend the deontologist's subjective value function to states of the world.
3. Disentangling thevalue ofconsequences (states of the world) and the value of actions is necessary in order to incorporate deontological reasons in the theory of rationality. At the same time it allows for solving some seeming paradoxes of rational choice theory, e.g. the Prisoner's Dilemma. The moral requirement to act cooperatively can now be fulfilled without falling short of optimizing subjective value. Acomprehensive description of strategies includes the structural traits of the game-format in which the strategy is embedded. Choosing the cooperative strategy in a Prisoner's Dilemma situation does not *ipso facto* violate Ramsey-coherence any more.
4. Regarding the class of deontologically motivated actions there is still a logical space for weighing processes of different kinds. A fully rational person chooses her actions such that she optimizes the state of the world within the constraints given by deontological rules she accepts. In some cases the disadvantages regarding the causal consequences for the state of the world may be so severe, however, that violations of deontological rules are justified. Deontological rules themselves have to be weighed against each other and probabilities of actions which result from my actions have to be taken into consideration, too.
5. Subjective value in this extended sense represents the whole of a rational person's reasoning process, including deontological reasons. *Extended subjective value* requires a *comprehensive description* of actions that allow for weighing consequentialist and deontological reasons alike. Extended subjective value correspondends adequately with the complexities of practical reasoning, whereas standard decision theory falls short of these complexities. Our teleological intuition is preserved insofar as rational action, including rational deontological action, optimizes subjective value.
6. An action is rational if it is motivated by adequate reasons, and if it is adequately related to these reasons. The rationality of action cannot be judged in isolation. Practical rationality is linked to the wider context of the acting person's beliefs and preferences. We expect that these beliefs and preferences meet some requirements of coherence.The extended axioms of decision theory are criteria for these requirements. A substantial theory of rationality is a theory about good reasons to act. Decision theory is a formal theory of rationality which defines some mini-

mal, but indispensable requirements of coherence. If these requirements are met, beliefs and preferences can be represented by two real-valued fonctions of subjective probability and subjective value. This numerical representation should not be confused with substantial rationality. Rational deontological action optimizes subjective value in the extended sense, that has been introduced in this article. The deontologist optimizes subjective value if her preferences are coherent. Since preferential coherence is a necessary but not sufficient, requirement of rationality, we know that the rational deontologist optimizes subjective value, but we do not know whether the deontologist who optimizes subjective value is rational.

7. If my argument is valid, decision theory is an all-embracing, but not a full-blown theory of practical rationality. Decision theory is *all-embracing* insofar as all rational actions comply with decision-theoretic criteria. It is, however, not a *full-blown* theory of practical rationality, since decision theory deals only with requirements of preferential coherence. Preferential coherence as it is defined by decision theory is a precondition of substantial rationality in general (deontological reasons included), it is not substantial rationality itself. A good theoryof practical rationalityshould be holistic and coherentist: A good theory of practical rationality should not exclude any essential element of the reasoning process from the start like deontological reasons (it should be holistic) and it should allow for a reasoning process covering all elements (it should be coherentist), no element should in principle be excluded from being weighed against other elements. Decision theory can indeed meet these requirements if it is based on *extended preferences* and *extended axioms of coherence* referring to *comprehensive descriptions of actions.*

Structural Rationality in Game Theory

Preliminary Remarks

When I began to think about normative questions of rationality and morality, I soon became a consequentialist. This is the natural result of questioning socially established institutional and moral rules and their metaphysical underpinnings. A critical mind cannot accept a duty to do this or that just because it is told that this duty is implied by some general principle. The immediate question is: Where does this principle come from and how is it justified? And if this question cannot be answered satisfactorily, some critical minds become normative sceptics and others search for rational grounds. Those who search for rational grounds easily end up with one form or other of consequentialism. A rational person does not act by established convention or acquired habit but chooses her action as a means to cause the preferred effects. Thus it seems that the rational person tries to optimize, and this seems to imply that moral action as a variant of rational action should optimize results from a universal point of view, e. g. by maximizing the sum of individual utilities.

In my case the story did not end there. When I became aware of the conflict between aspects of distributive justice and this form of consequentialism, I had second thoughts about moral reasoning in general. Questions of justice have at least equal weight as questions of utility-efficiency. There is a conflict between the simple extension of individual rationality to social or moral rationality on the one hand and the central moral intuition which does not allow us to disregard aspects of distributive justice. But if one takes this argument seriously, the door is open to include further considerations of different moral intuitions. The initial effort to rationalize normative discourse eventually resulted in dismissing almost all of our normative convictions, which is the dead end of normative theory. A good normative theory has to be compatible at least with the most central and immanently coherent normative convictions we share. This is the basis of the critique of consequentialism not only as an ethical but also as a theory of practical rationality (cf. Nida-Rümelin 1993, 1995).

The critique of consequentialism has shown that normative evaluation cannot be external: what we value as good or bad, right or wrong, depends on a complex structure of good reasons for actions. These reasons can only partly be reconstructed by the consequentialist device. To act irrationally is to act so that the balance of good reasons is not in favor of that act. The accepted reasons represent a form of social community, i.e. they constitute social interactions as legitimate or illegitimate. Likewise the person in choosing her favorable good reasons for action chooses an individual form of life. To act irrationally and to act immorally means not to comply with these (normatively acceptable) interpersonal and intrapersonal structures.

If the divergence between normative and empirical analysis is endemic to all social analysis, one should have second thoughts about the adequacy of the normative theory. There can be no doubt that there are systematic misrepresentations of

Published in W. Leinfellner and E. Köhler eds., *Game Theory, Experience, Rationality – Foundations of Social Science, Economics and Ethics*, Kluwer, Dordrecht, 1997.

decision situations caused by our limited capacity to deal with complex strategic situations and probability distributions, but there are simple cases where this kind of reconciliation is not convincing. The notion of structural rationality can help to integrate empirical and theoretical findings especially in game-theoretic analysis into a coherent conception of practical rationality instead of dealing with these findings as a threat to or a subversion of rationality (cf. Elster 1979, 1983, 1987).

The notion of structural rationality implies only a moderate theoretical change. It gives up consequentialism as the standard interpretation of decision theoretic rationality. But coherentism is preserved insofar as the overall preferences of a rational person can still be represented by a real-value function (cf. Nida-Rümelin, 1997 chap. VI). In more philosophical terms: starting from a Humean account of motivation, one is forced to acknowledge structural constraints on optimization in order to define practical rationality. And likewise: starting from Kantian apriorism as established by the universalizability test of the *Categorical Imperative,* one has to refer to contingent personal preferences in order to decide what is rational and moral. The notion of structural rationality in that sense covers a middle-ground between orthodox Humeanism and orthodox Kantianism. Humeans should acknowledge that rational persons strive inter alia at inter-temporal consistency (or better: coherence), and this constrains consequentialist optimization, and Kantians should acknowledge that without contingent" personal preferences (*Neigungen*) the normatively required action (moral if rational in the Kantian sense) would be vastly underdetermined. Structural rationality is incompatible with both Kantian apriorism and Humean consequentialism.

The basic idea of structural rationality is the following: if we decide in concrete situations between specific options, we take into consideration "structural" aspects of our plans of life (to use a Rawlsian term which is easily misinterpreted, as if we choose at a given time once and for all a specific personal plan of life), e.g. we have a preference for participating in a specific social project constituted by many actions over a long period of time by many participants and decide in a concrete situation for such an option which is at least compatible with our further participation or is even part of the overall social activity fostering that project. In cases like these we do not look only at the (causal or probabilistic) consequences of this concrete optional alternatives, instead we decide on the basis of more or less intuitive criteria of fit or compatibility. There is an analogy between the interpersonal and the intrapersonal case. In cooperating with others, we refrain from optimizing subjective values in order to be structurally coherent and reliable for others, and in case of personal projects, we refrain from optimizing parts of our life in order to make our life on the whole coherent. This sketchy characterization is everything but precise and explicit, and we will not try to give a general account of structural rationality here. Instead we pick up some problems of rational decision in game theory which can well be embedded into the notion of structural rationality.

Gauthier's Dispositions

Following Gauthier, the just person is disposed to comply with the requirements of the principle of minimax relative concession in interacting with those other fellows whom he believes to be similarly disposed. There are strong arguments showing that

there is no unique solution for rational bargainers as Gauthier assumed, but this criticism is not essential for our argument. Let us assume that Gauthier succeeds and there is one and only one solution for rational bargainers, which is the principle of minimax relative concession. Then we would face again the seemingly paradoxical result that the rational person would decide, "on utility-maximizing grounds, not to make further choices on those grounds" (Gauthier 1986, p.158).

Now the question is: what interpretation is needed to make this claim a sound one? Obviously it would be a sound claim if specific options were available to the agent. Think of the following options: there are ten pills, swallowing one of these pills establishes specific dispositions in the behavior of the agent who took the pill. If among these pills there is one which establishes a disposition to behave in compliance with the principle of minimax relative concession and another pill would establish the disposition to maximize personal utility (to discriminate this kind of maximization, we call it "pointwise maximization"), then the rational utility-maximizing agent would indeed prefer to take the first pill.

The problem with Gauthier s theory is that he does not offer such an interpretation. Instead he gives the impression that, if to behave according to a specific disposition had better results than pointwise maximizing behavior, then rational maximization would require us to act according to this disposition. In this way, Gauthier tries to reconcile two forceful intuitions (1) that rational action is optimizing and (2) that constraints are necessary in order to act rationally (in this case to cooperate). This reconciliation hides the breakdown of pointwise optimization. What Gauthier and many other practical philosophers and decision theorists have discovered is that the optimizing theory of rationality is indeed self-defeating. My proposal here is to face these results and modify the theory with the help of the notion of structural rationality.

In the case of Gauthier, the interpretation is straightforward: the rational agent wants to cooperate. If the principle of minimax relative concession could indeed uniquely establish social cooperation – as Gauthier argues –, then agents wishing to cooperate would indeed comply with this principle. A person complying with this principle would therefore reveal a certain "disposition" in acting and interacting and thus comply with a specific social structure of interaction. But since the person did never choose dispositions – instead she chose concrete actions complying with the principle –, she in fact did not optimize in all these singular cases: which is compatible with Gauthier's idea that the rational person is an overall optimizer of her (social) life. It might well be that the person is doing both in the singular case: optimizing and complying, but this is a mere contingent combination.

The rational person opts for a structural trait of her (inter-)action because she knows that general compliance with this structural trait would optimize her interests or (the weaker claim) because only in complying can she find other persons willing to cooperate – but complying means not to optimize in the singular (general) case.

Multi-Person Cooperation

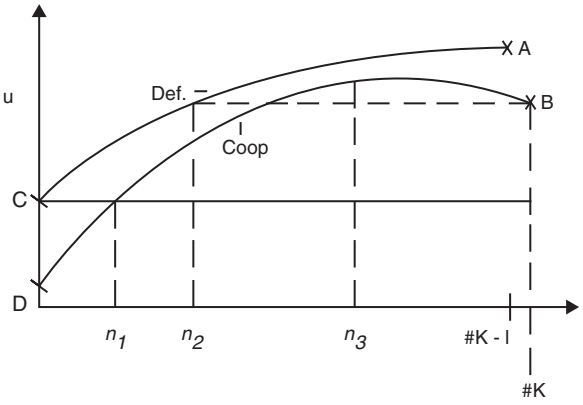
Let us assume that, depending on the number of cooperators, the individual utility status changes. Let us further assume that the individual utility status of a person is always lowered if this person decides to cooperate instead of not to cooperate, if the

rest of cooperators and non-cooperators remain unchanged. Therefore we assume that there is an individual incentive not to cooperate for every person involved. A closer analysis of the multi-person cooperation problem (cf. Schelling 1978, chap. 7) shows that there are some differentiations possible which possibly have an impact on the motivation to cooperate.

For example it could be of some relevance for the motivation of rational agents whether an increase of cooperators would increase or decrease the individual utility status of all the cooperators. In some cooperation problems, there is an optimal number of cooperators (n_3) beyond which there is no cooperative surplus (see illustration below).

It might also influence the motivation of rational persons whether the cooperative surplus suffices to get the utility statuses beyond the level of universal noncooperation. One might call the number of cooperators necessary for getting a utility status of the cooperators which is higher than the utility status in case of universal non-cooperation the "threshold for self-sustaining cooperation" (n_1). It is not unreasonable to assume that below this threshold the motivation to cooperate is difficult to obtain. If there are more than n_1 potential cooperators, one could expect that they base their cooperation on an agreement to cooperate even if this agreement is not enforced. A person knowing that there are less than n_1 cooperators will probably feel that personal cooperation is senseless.

In general, free-riders profit from the cooperation of others: they are better off if others cooperate. We assume that their utility status rises monotonously with the number of cooperators. Beyond a number of cooperators (n_2) the non-cooperator might be better off than the cooperator in case of universal cooperation.



A: Unilateral Non-Cooperation
B: Universal Cooperation
C: Universal Non-Cooperation
D: Unilateral Cooperation
$u(A) > u(B) > u(C) > u(D)$
n_1: threshold for self-sustaining cooperation
n_2: threshold for free-riding improvements to universal cooperation

n_3: co-ordination optimum for cooperators
#K–I: cooperation optimum for non-cooperation
#K: cooperation maximum

Within the standard consequentialist account of rational action all participants would end up in C. But in many everyday situations in which multi-person cooperation is possible, individuals are motivated differently, i.e. under specific conditions they are willing to cooperate even if the consequences of cooperation in terms of their individual utility status are such as described in the illustration. Agents of this kind typically do not cooperate unconditionally. They cooperate because they see their individual action as part of a multi-person collective action. They are in favor of this collective action and therefore do their individual part in this collective action. Since in many cases such a collective (cooperative) action is not constituted by individual actions of the same specific type and in others there might be different possible collective actions which constitute cooperation, there are difficult problems how to make this kind of "belonging" of individual actions as part of a collective action explicit (cf. Tuomela 1995). The situation as illustrated above Can therefore be quite complex in detail. A simple version would result e.g. if there were at least n_1 persons being motivated to cooperate under the condition that at least $n_1 - 1$ others are willing to cooperate, too. The individual person then chooses her cooperative action as part of a collective action comprising at least n_1 actors.

Metagames and Structural Rationality

The origin of metagame theory goes back to von Neumann and Morgenstern (1953). They expressed the view that in order to analyze two-person zero-sum games in normal form one should analyze the so-called minorant and majorant games, i.e. the games which would result if one of the players chose her strategy in knowledge of the others' choice.

In more general terms, Howard (1971) introduced the notion of a metagame that results for every game in normal form if one of the interacting persons chose her strategy in knowledge of all the others' choices.

In order to construct a metagame one makes use of the fact that to every given arbitrary normal form of a game there corresponds at least one extensive form (assume that the persons choose in some fixed order and that they are not informed about any strategy choices preceding their own, which results in quite big information sets containing all decision nodes of the person).

The extensive form of a metagame for person *i* results if we develop the extensive form out of the normal form of a game as above with *i* coming last and then delete *i's* information set and give her one information set for each of her decision nodes: she chooses last knowing the others' choices. Thus there are *n* metagames in extensive form resulting out of one *n*-person game in normal form.

The normal form of the metagame is found if we replace *i*'s strategy set with a set containing all functions from the joint strategy choices of the others to his own strategy choices.

Mixed strategies are excluded by the non-quantitative approach of metagame theory.

Original PD:

		B	
		C	D
A	C	3/3	2/2
	D	4/1	2/2

B – PD:

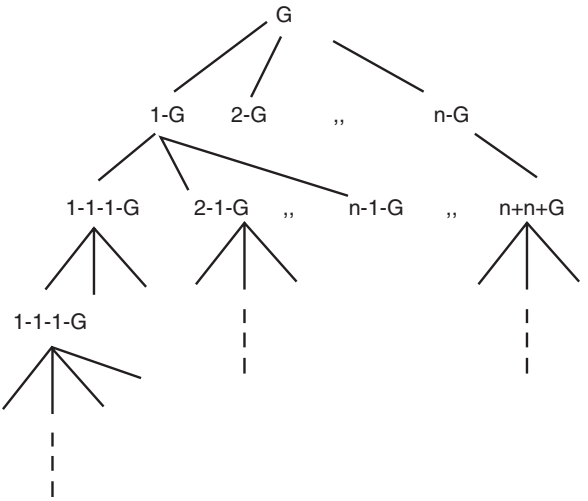
f_1: C → C & D → C
f_2: C → C & D → D
f_3: C → D & D → C
f_4: C → D & D → D

		A	
		C	D
B	C & C	3/3	1/4
	C & D	3/3	2/2
	D & C	4/1	1/4
	D & D	4/1	2/2

Equilibrium in: D & D/D

On the basis of the simple Prisoners' Dilemma there results a PD-metagame which reveals in addition to universal non-cooperation universal cooperation as an equilibrium point (which is not lost in the further extension of the metagame tree).

Game → i-Metagame
i-Metagame → ji-Metametagame
→ Infinite Metagametree



	CC	DD	CD	DC
C C C C	33	14	33	14
D D D D	41	22*	22	41
D D D C	41	22	22	14
D D C D	41	22	33*	41
D D C C	41	22	33	14
D C D D	41	14	22	41
D C D C	41	14	22	14
D C C D	41	14	33	41
D C C C	41	14	33	14
C D D D	33	22	22	41
C D D C	33	22	22	14
C D C D	33	22	33*	41
C D C C	33	22	33	14
C C D D	33	14	22	41
C C D C	33	14	22	14
C C C D	33	14	33	41

Three equilibrium points:

1. DDDD/DD
2. DDCD/CD (pareto-efficient)
3. CDCD/CD (pareto-efficient)

The problem with this metagame "solution" of the Prisoners' Dilemma is that there is no immediate and straightforward interpretation of meta-strategies. The standard conviction in game theory is that the metagame approach should be given up in favor of the supergame (iterated game)-analysis. Indeed some of the most interesting results of metagame theory can be reproduced in supergame analysis. Instead of reformulating the decision problem diachronically (supergames), structural rationality allows for understanding metastrategies as a reconstruction of some motivational process. The above example of the PD-metagame is a good illustration:

(a) CC can be interpreted as: "I cooperate in both cases (case 1: I expect the other person to cooperate, case 2: I expect the other person not to cooperate)."
(b) CD can be interpreted as: "I cooperate if I expect the other person to cooperate and I do not cooperate if I expect the other person not to cooperate."
(c) DDCD can be interpreted as: " I do not cooperate if I expect the other person to be a universal cooperator (CC); I do not cooperate if I expect the other person to be a universal non-cooperator (DD); I do cooperate if I expect the other person to be a conditional cooperator (cooperating only if she expects the other person to cooperate too); I do not cooperate if I expect the other person not to cooperate if she expects me to cooperate, and to cooperate if she expects me not to cooperate."

The new equilibrium point DDCD/CD can now be interpreted synchronically: Neither of the two prisoners has an incentive to change the decision if these decisions are motivated like (c) for prisoner A and (b) for prisoner B. This motivational structure makes sense only if both prisoners see their individual decision as constitutive part of a (fictitious) structure of interaction, partly reconstructed as metastrategies.

Rational Inducement

In cases of rational inducement there are several equilibrium points with different outcomes for the participants. In such situations a person who is willing to accept a disadvantage can threaten another participant successfully if the other participant is a simple optimizer. Only in situations of complete transparency can one say that the actual non-optimizer succeeds against an optimizer in getting his preferred equilibrium.

Analogously with meta-strategies as discussed above, there is a diachronic interpretation in terms of supergame analysis and a synchronic interpretation in terms of structural rationality.

		B	
		b_1	b_2
A	a_1	3/4	3/3
	a_2	1/1	4/3

Two Equilibria: a_1/b_1 and a_2/b_2

$\min(a_1) = 3 \quad \min(b_1) = 1$
$\min(a_2) = 1 \quad \min(b_2) = 3$

A prefers the equilibrium a_2/b_2 ; *B* prefers the equilibrium a_1/b_1. If both were consequentialist optizimers, the chosen equilibrium point would be underdetermined. And if there would be no possibility to coordinate, they could end up with any of the four possible strategy combinations. The moment in which one participant informs the other which strategy he will choose, the other is (as a consequentialist actor) bound to accept the "proposed" equilibrium point. The conflict is "solved" by time precedence. But if one of the participants (*A*) is a structurally rational agent, choosing his decisions as part of bigger (structural) strategies, he will possibly opt for a_2 unconditionally. Thus accepting the possible disadvantage of getting 1 instead of 3, if *B* chooses b_1. But if *B* is a consequentialist actor, the underdetermination is resolved because *B* would give in and stay to b_2.

The Utilitarian Solution to the Cooperation Problem

As long as there are interpersonally variant overall action-guiding value-functions, it is to be expected that the cooperation problem persists. In theory though, it is possible to constrain the sets of feasible (interpersonally variant) value functions such that the cooperation problem does not arise. But if we assume – as utilitarianism in general does – that there are two levels of values – the values of the second level being an aggregate of the values of the first –, then we see that under plausible conditions there is only one kind of aggregation; which eliminates the cooperation problem. If we constitute the action guiding moral preferences of the ideally rational and moral person by transfer rates, which adds a portion of the first level-values of other persons, then the only transfer rate which puts an end to the cooperation problem is 1 (100%). If the interpersonal transfer rate is lower, some PD-situations on the first level will survive on the second, and if it is higher, then some non-PD-situations on the first level will transform into a PD-situation on the second. This result can be taken as an argument in favor of utilitarianism, or it can be taken as an argument in favor of a deontological theory of constraints. In the first case one concentrates on withdrawing the cooperation problem and concludes that only the utilitarian constitution of action-guiding moral preferences will have this result and therefore take utilitarianism to be well-founded. In the second case one concludes that persons can cooperate even if cooperation is not individually optimizing.

I am convinced that only the second conclusion can be defended (cf. Nida-Rümelin 1991). But even if one prefers the first conclusion, the case for structural rationality as the adequate normative account would – astonishingly enough – not be lost. Even if we accepted the somewhat radical idea that ideally rational and moral persons share the same action guiding preference relation, there would remain a problem of cooperation which is responsible for the superiority of ruleutilitarianism (if adequately reconstructed and under some specification of application).

The collective irrationality of optimizing agents with interpersonally different utility functions in the generalized PD can be solved by a radical measure. The measure is to eliminate interpersonal differences on the level of action guiding preferences. Utilitarianism is one way to do this. Preference utilitarianism as a special case provides one interpersonally invariant value function by simple aggregation of preferences. The most sophisticated form as presented by Harsanyi, combines decision theoretic Bayesianism with Preference Utilitarianism, it aggregates personal preferences to moral ones. Moral preferences can then be represented by an interpersonally invariant value function.

One might think that in this way the collective irrationality of optimizing behavior has finally ceased (even if some might think that the price for achieving this result is too high). But this is not true. The simple reason is that even rational ethical consequentialists face problems of cooperation. Problems of cooperation transformed into mere coordination problems survive in the ideal world of optimizing utilitarians (having the same interpersonally moral value function); therefore even rational utilitarians ought to change from consequentialist rationality to structural rationality.

Consequentialist utilitarians usually deny this. The most prominent form of this denial is the pretended practical equivalence of act- and rule-utilitarianism. The "proofs" of this thesis were refuted soon (cf. Ezorsky 1968) after they had been presented (cf. Lyons 1965), but the thesis surprisingly survived and is still one of the standard arguments in favor of ethical consequentialism.

Intuitively it is quite simple to understand why the equivalence thesis must be wrong. The argument in favor of the equivalence thesis is that the possible rules should not be taken out of a given list – given by language or conventional rationality – but can be formulated in as complex, detailed and sophisticated a way as you want. Even if there are pragmatic limits for this refinement, in theory all the anti-consequentialist arguments can be met by refining the respective rule. If the conventional rule of truth telling has good utilitarian consequences in general, then one should include the exception in a modification of the rule, and *a limine* you end up with the consequentialist rule to tell the truth if and only if it (this concrete case of telling the truth under these concrete circumstances) has optimal consequences. Rules survive, but only as rules of thumb. Those who think that even in theory (i.e. in the ideal case of perfectly rational and moral agents) one should follow rules also when this rule-following does not have optimal consequences, attribute an intrinsic moral value to rules, and this cannot be accepted within utilitarian approach.

The problem with this argument is that it is based on a parametric, not a strategic (or game-theoretic) view regarding practical rationality. Even utilitarians face the

problem of how to coordinate their actions interpersonally. If we look at the matter from a game-theoretic point of view, we easily see that optimizing a collective strategy (i.e. a combination of individual strategies) does not have the same result as optimizing an individual strategy except for rare special cases like the ideal market. The equivalence thesis is therefore wrong. It has to be conceded though that proponents and critics of the equivalence theses did not realize that the equivalence thesis can be read in two different ways: parametrically and strategically. Strategically it is obviously false. Parametrically its falsity is less easily demonstrated and we do not need to do so here.

Harsanyi has proven the non-equivalence of act- and rule-utilitarianism in analyzing some examples of coordinative collective action by game-theoretic tools (cf. Harsanyi 1977, 1983, 1985). Since the argument is straightforward and without doubt, it remains only to interpret the rule-utilitarian device adequately. Rule-utilitarians in Harsanyi's sense choose that action which performed by all other moral and rational (i.e. rule-utilitarian) persons maximizes overall expected preference fulfilment. This means that in the general case the rational and moral person does not optimize the (preference-) utilitarian value function, but that the collective of rule-utilitarian agents jointly optimize the (preference-)utilitarian value function. Thus the rational moral person chooses her action as part of a collective strategy which is optimizing, whereas the individual action is not optimizing. The rational and moral person acts structurally rational.

References

Elster, Jon. 1979. *Ulysses and the Sirens.* Cambridge: Cambridge University Press.

Elster, Jon. 1983. *Sour Grapes.* Cambridge: Cambridge University Press.

Elster, Jon. 1987. *Subversion der Rationalität.* Frankfurt a.M./New York: Campus.

Ezorsky, Gertrude. 1968. A Defense of Rule Utilitarianism. *Journal of Philosophy* 65: 533–544.

Gauthier, David. 1956. *Morals by Agreement.* Oxford: Clarendon Press.

Harsanyi, J.C. 1977. Rule Utilitarianism and Decision Theory. *Erkenntnis* 11: 25–53.

Harsanyi, J.C. 1983. Basic Moral Decision and Alternative Concepts of Rationality. *Social Theory and Practice* 9: 231–244.

Harsanyi, J.C. 1985. Does Reason Tell Us What Moral Code to Follow and, Indeed, to Follow any Moral Code at all? *Ethics* 96: 42–55.

Howard, Nigel. 1971. *Paradoxes of Rationality.* Cambridge, MA: MIT Press.

Lyons, David. 1965. *Forms and Limits of Utilitarianism.* New York/Oxford: Clarendon Press.

Nida-Rümeiin, Julian. 1991. Practical Reason or Metapreferences? An Undogmatic Defense of Kantian Morality. *Theory and Decision* 30: 133–162.

Nida-Rümelin, Julian. 1993/1995. *Kritik des Konsequentialismus.* München: Oldenbourg.

Nida-Rümelin, Julian. 1997. *Economic Rationality and Practical Reason.* Dordrecht: Kluwer [in print].

Schelling, Thomas C. 1978. *Micromotives and Macrobehavior.* New York: Norton.

Selten, Reinhard. 1978. The Chain Store Paradox. *Theory and Decision* 9: 127–159.

Sen, Amartya. 1976/1977. Rational Fools: A Critique of the Behavioural Functions of Economic Theory. *Philosophy and Public Affairs* 6: 317–344.

Tuomela, Raimo.1995. *The Importance of Us. A Philosophical Study of Basic Social Motions.* Stanford: Stanford University Press.

von Neumann, John, and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior.* Princeton: Princeton University Press.

Rationality: Coherence and Structure[25]

Abstract
In this paper, it is argued that the standard decision theoretic axioms provide core principles action-guiding preferences of rational agents should be required to satisfy. This, however, does not involve a commitment to consequentialism.

It can be rational to be polite. To behave politely requires to follow certain rules. A. Sen has shown that following these rules is incompatible with acting as a consequentialist. Sen proposes, therefore, to give up some principles of decision theory. I argue instead that a 'comprehensive' description allows for both: being polite *and* acting in accordance with the principles of decision theory. This is possible because these principles should be interpreted as an expression of coherentism, not of consequentialism.

Introduction

The ideal rational person has coherent preferences, i.e., preferences satisfying the decision-theoretic axioms of rational choice. A person with, say, intransitive preferences cannot be perfectly rational. A similar statement seems to be true for every single axiom out of the set of requirements constituting the theory of expected utility maximization. This theory is a weak theory of rationality. It is Unable to discriminate between good and bad reasons for acting. It only qualifies revealed preferences as coherent or incoherent.

Revealed preferences are the result of practical reasoning. In the last resort, rationality is constituted by actions guided by good reasons. Therefore, we do not know whether a person is rational if we know that the person maximizes expected utility. If a person does not maximize expected utility we know that she has incoherent preferences and is irrational.

Let us call this view the *compatibility assumption:* Whatever are the reasons which guide a person's actions, action-guiding rational preferences fulfill the standard axioms of rational choice if the person is to be considered rational. Rational agency must be compatible with a description under which the person maximizes expected utility.

The intuitive underpinning of the compatibility assumption is not consequentialist but strictly coherentist. It is not based on the idea that rational action means to optimize consequences, because there are many types of good reasons for action which cannot be reconstructed within a consequentialist theory of rationality.[26] Certainly, this is only a prima facie argument against consequentialism. It does not prove that consequentialism is wrong. Since there are good prima facie reasons against consequentialism, consequentialism could be taken to be true only as a

Published in J. Nida-Rümelin and W. Spohn, Rationality, Rules, and Structure, Kluwer, Dordrecht, 2000, pp. 1–15.

result of some convincing theory. In this paper, I do not exclude the possibility of such a convincing theory. What I take to be granted is that there are prima facie arguments against consequentialism such that consequentialism cannot be the starting point of the argument.

Two examples might suffice for rendering the assumption plausible that many of our good reasons for action – moral and non-moral ones[27] – are non-consequential.

First Example The fact that Tom has promised (at time t) to come can be a good reason for Tom to come at some later time t'. It is not necessary to spell out the conditions under which a given promise in fact *is* a good reason to keep it in order to see that these conditions cannot be confined to considerations regarding outcomes alone. If Tom keeps his promises only in case he expects the outcome to be optimal, he has not really understood what it means to give (and to keep) a promise. If Suzan knows that Tom keeps his promises only in case he expects the outcomes to be optimal she will not trust him. In this case, promise-giving would lose its power to coordinate intentions and actions. If we assume that forming the intention to keep the promise is part of the act of (genuine) promise-giving – as John Austin assumed in his speech-act analysis[28] – then a consequentialist intention would even be *incompatible* with giving a promise. This does not exclude that certain or probable consequences of keeping the promise vs. not keeping it are relevant for determining whether Tom has a good reason to come. It does exclude, however, that the outcomes of keeping the promise vs. not keeping it are the only aspect relevant for determining whether Tom has a good reason to come.[29]

Second Example Tom might have the intention to cooperate in a specific situation with Suzie, expecting that Suzie has the same intention.[30] Let us assume that cooperation can be defined with regard to a matrix of outcomes which shows the structure of a Prisoner's Dilemma. Tom might intend to cooperate although he is aware of the outcome structure of the game he plays. It is not necessary that Tom attributes some kind of intrinsic value to the act of cooperation itself. He might deliberately choose a dominated strategy in cooperating. In this case Tom would not maximize the expected value of outcomes. Nevertheless, Tom could have coherent preferences. The intention to cooperate should be compatible with having coherent preferences. According to the definition above a coherent preference relation satisfies the decision-theoretic axioms. It should, therefore, be conceptually possible that Tom (in cooperating) maximizes expected utility, although he chooses (in cooperating) the dominated action regarding outcomes. This might seem strange at first sight, but it leads to the core of an adequate analysis of rational action which drives a wedge between consequentialism and coherentism.

The observation that there are good reasons for action which are not consequentialist, i.e., which recommend actions which do not maximize the expected values of their outcomes, can lead to quite different reactions:

(1) One might assume that those prima facie reasons which cannot be integrated are no genuine good reasons (this account could be called 'consequentialist'). The

problem with this account is that some of our most central types of reasons would have to be excommunicated.

(2) One might confine the range of application of rational choice and exclude moral and other types of good reasons for action (this could be called the 'narrow account of rationality'). Rational choice theory would then not be an all-embracing theory of practical rationality any more.

(3) One might give up some of the axioms constitutive for standard rational choice theory as it is done e.g., by McClennen (1990) in his theory of resolute choice (this account could be called 'revisionist').

(4) One might redesign the conceptual framework, i.e., reinterpret the basic concepts of rational choice such that non-consequentialist reasons for action can be integrated (I call this the account of 'structural rationality' for reasons which I hope to become clearer later on).

The aim of this paper is to provide some arguments lending support to the fourth option.

Weak Coherentism

I shall call a preference relation 'weakly coherent' or 'Ramsey-coherent' if it satisfies the standard decision-theoretic axioms. This is a 'weak' requirement since although the rationality of an action depends on a complex framework of conative and epistemic attitudes, this framework is not to be considered when testing whether a preference relation satisfies the decision-theoretic axioms. An action is said to be rational *in the full sense* if, additionally, the action and the framework are coherent.

Preferences which do not satisfy the standard decision-theoretic axioms are not coherent. However, preferences which do satisfy the standard decision-theoretic axioms can still be incoherent in relation to the broader frame of conative and epistemic attitudes. Preferences which satisfy the standard decision-theoretic axioms are weakly coherent. Taken together, the standard decision-theoretic axioms are necessary and sufficient criteria for weak coherence, but they are only necessary criteria for coherence in the broader sense.

We can think of many different types of conative attitudes (e.g., desires, wishes, hopes, intentions, preferences) and of many different theoretical possibilities for systematizing this plurality. It seems, however, that, independently from the type of the respective conative attitudes and their systematization, there is an initial plausibility of (weak) coherentism or Ramsey-coherentism, because (and insofar as) conative attitudes lead to singular actions. There might occur irresolvable conflicts between different desires, hopes and wishes, there might even be genuine moral dilemmas, but in the end we have to integrate our conative attitudes such that we are able to act. If, for example, our action-guiding (overall) preferences were incomplete, then in some situations we could not act, or, to put this in a différent way: our actions would lose their basis in our conative attitudes. The violation of each singu-

lar axiom would have the same devastating result for the agency of individuals trying to decide rationally. Preferences – however they may be constituted – have to be Ramsey-coherent if they can be thought to guide rational action.

Irrationality of action-guiding preferences has two main sources, depending on

(1) *how* they are based on conative (and epistemic) attitudes and
(2) on *which* conative (and epistemic) attitudes they are based.

Even if the action-guiding preferences cannot be criticized for being inadequately based on conative attitudes or for being determined by inadequate conative attitudes (because the determining conative attitudes are inadequate), they can still be criticized if they are (weakly) incoherent.[31]

Since Ramsey-coherent preferences can be represented by a real-valued utility function (determined up to positive linear transformation), quantitative representability (in this sense) is a minimal requirement for ideal rationality. This explains the use of the term 'compatibility assumption' which I have introduced above: the plurality of our reasons for actions has to be made coherent such that the action-guiding overall preferences conform with the axioms of rational choice. If they conform we do not know whether these preferences are rational, but if they do not conform we know that at least one of these preferences is irrational. Compatibility is a relation holding (or not holding) between reasons for actions on the one hand and coherence of action-guiding preferences on the other. If our reasons for actions are such that they are incompatible with Ramsey-coherent action-guiding preferences, they have to be modified, whatever their content may be.[32] It cannot, e.g., be rational to have action-guiding (overall) preferences which are not transitive, No type of reason for action can be held responsible for violating reflexivity, completeness, transitivity, continuity, reduction, substitution or monotonicity.[33]

Some familiar arguments in favour of the rational choice axioms like the money pump argument or the dutch book argument might inhibit a proper understanding of my argument. I postulate an initial plausibility of the axioms of rational choice as minimal coherence requirements of action-guiding preferences, whereas the traditional defenders of the axioms of rational choice use consequentialist arguments. However, since many of our good reasons for actions are not consequentialist we should not rely on consequentialist, but on coherentist intuitions. Our coherentist intuitions do favour the compatibility assumption. There is a prima facie justification of the rational choice axioms which does not depend on consequentialism.

Although standard arguments in favour of the axioms of rational choice are based on consequentialist intuitions, there is a kind of critique of the axioms of rational choice which also is consequentialist in spirit. If an agent, who maximizes expected utility, is in the end worse off than an agent who does not maximize, conseqentialism seems to require to disregard the axioms of rational choice. To avoid confusions, it is helpful to distinguish terminologically between *general consequentialism* as the view that consequences and their value determine exclusively what is rational and the more specific view that an individual action is rational if and only if its consequences are optimal – let us call this latter view *agency consequentialism.*

At first sight, it seems that this distinction does not make much sense. Agency consequentialism obviously is a specialization of general consequentialism. It seems that a general consequentialist must be agency consequentialist, too. The reversal, though, does not hold. If one adopts the plausible stance that only concrete actions of natural individuals can bear the predicate 'rational', the consequentialist account had to be confined to agency consequentialism anyway. Every different usage would on this account be at its best acceptable as merely metaphorical. However, even if there are other entities to which the predicate 'rational' can be applied (e.g., plans of life), the rationality of these entities is not implied by the rationality of singular actions.[34]

The distinction between agency consequentialism and general consequentialism is useful to understand self-defeating arguments against general consequentialism. Tom, who judges the course of his life up to now using the time integral of some value function which – say – attributes to his mental state at every point of time a real number representing its subjective quality, realizes that he would have done much better if he had refrained from choosing every singular action such that it optimized its causal (including probabilistic) consequences. If he is right – and I suppose he is –, consequentialism applied to whole courses of life would be incompatible with agency consequentialism; general consequentialism would be self-defeating.

All forms of consequentialism assume that there is no intrinsic value of rules or types of actions and that in the last resort it is utility or preference fulfillment which renders an action rational or not. In this weaker sense one can be consequentialism without being in the stricter sense consequentialist, i.e., agency-consequentialist. A person, acting rationally in the sense of agency-consequentialism, can be overall irrational regarding overall individual preference fulfillment. This difference is used, e.g., by Edward McClennen in his theory of resolute choice which turns general consequentialist arguments against the specific form of agency-consequentialism. I agree with McClennen in maintaining that practical rationality is not the result of pointwise ('myopic') maximization: agency consequentialism is no convincing general theory of practical rationality. But (as opposed to McClennen) I am convinced that this does not force us to give up the theoretical core of decision theory, if only we rely on a strictly coherentist interpretation (or foundation) of its axioms.

The observation that there are rational actions which do not maximize the expected value of their outcomes leads McClennen to give up axioms constitutive for standard rational choice theory. My proposal, against this move of McClennen, is to redesign the conceptual framework, i.e., to reinterpret the basic concepts of rational choice such that non-consequentialist reasons for action can be integrated (this is the point of the *compatibility assumption*). The theory of resolute choice favours the third reaction of the different options mentioned above, whereas I think we should opt for the fourth alternative.

The justification of the compatibility assumption consists of two elements. The first is to stress its initial plausibility and the second is to show that arguments against it are not convincing. In a recent illuminating article Sen[35] has developed arguments which, if they were stringent, would undermine the initial plausibility of

the compatibility assumption. In order to refute them as arguments against the compatibility assumption I delineate an alternative interpretation of the rational choice framework. I hope that in doing so it will become transparent how arguments of this type in general are to be answered.[36]

Comprehensive Description

Let us sketch the idea of a 'comprehensive' description of decisions using one of Sen's examples: you arrive at a garden party and can readily identify the most comfortable chair. You would be delighted if an imperious host were to assign you to that chair. However, if the matter is left to your own choice, you may refuse to rush to the chair. You select a 'less prefered' one. The reason for this behaviour is that you want to be polite and that you take it as a politeness-constituting-rule not to take the most comfortable chair (if there is only one available) in such situations. It is not the case that you merely want to give the impression to be polite, neither do you fear any bad consequences if you were not polite – you just want to be polite. To want to be polite cannot be excluded as a conative attitute which may become reason-giving in adequate circumstances. To aim at being polite cannot be charged with being an irrational attitude, I suppose. This conative attitude is determined by accepting (and finally acting in accordance with) the very rules constituting politeness.

In this example there seems to be a chooser dependence involved which we cannot get rid of in passing from self-oriented (egotist) preferences to all-comprising ones (i.e., consequentialist) which regard the whole situation including the distribution of chairs to all the other guests since we additionally assumed that you prefer to have the most comfortable chair if you get it as a result of another persons choice (the imperious host).

To accomodate preferences of this kind it is necessary to take the act of choice itself into consideration. In this example it is obviously relevant who chooses even if the outcomes remain unchanged. A comprehensive description of the situation should, therefore, include the information who chooses. Such a description cannot be integrated into consequentialism, because from a consequentialist point of view it is the subjective value of consequences alone (i.e., the state of affairs as it results from the respective choice) which determines rational choice. If polite behaviour can be rational and if polite behaviour implies chooser dependence, then it cannot be described by reference to consequences (in the technical sense described) alone. The rationality of politeness requires a non-consequentialist conceptual framework: we have to know more about the options than merely their consequences. If actions are described exclusively by reference to lotteries which have genuine consequences as outcomes, it is impossible to integrate chooser dependence into the rational choice framework. And if it is impossible to integrate chooser dependence, it is impossible to characterize this kind of polite behaviour as rational. Since rational agents ought to be polite like any other agents, the theory of rational choice should be modified such that *conceptually* they can be polite (this is a peculiar case of OUGHT implies CAN).

Since there is no justification for the assumption that having chooser dependent preferences implies having incoherent preferences, this modification should not lead to abandoning the core of rational choice theory. An adequate theory of rational choice should allow for preferences of this kind to be included (i.e., chooser dependent preferences). We should not accept a theory which implies that chooser dependent preferences necessarily lead to a preference relation which is Ramsey-incoherent. Chooser dependent preferences have to be described such that the corresponding subjective preference relation of the rational person is coherent. A strictly coherentist interpretation of the decision-theoretic conceptual framework should be compatible with all kinds of reasons. It is not the content (the motive on which a preference is founded), but the form in which it relates to other preferences which is relevant for deciding whether it can be integrated into a coherent preference relation. To have chooser dependent preferences, which is one kind of rule sensitive preferences, does not necessarily mean to have incoherent preferences.

In what follows I shall clarify some formal implications: let X be the set of alternatives where each alternative is characterized by the respective chair you are seated on. Then it seems that the principle of revealed preference is violated in our example: you have a preference for the state of affairs in which you sit on the most comfortable chair, but you choose to sit on another chair.

The principle of revealed preference can be stated as follows:

$$\forall x, y \in X : x, y <\in g(i) \leftrightarrow \exists X^{\circ}\left[\subseteq X \wedge x \in C\left(X^{\circ}\right) \wedge y \in X^{\circ}\right]$$

g: preference structure
$g(i)$*:* preference ordering of individual i
$i \in K$*:* the individual out of the collective K
X*:* optional set, set of alternatives
C: choice function (representing the agent's actual choices)

Let us take a new situation into account in which another equally comfortable chair is added. In this situation you could choose one of the two most comfortable chairs knowing that there is still another equally comfortable available. With this choice two quite basic conditions of rationality – the principle of contraction coherence α and the principle of conjunction coherence β – seem to be violated. Formally, these principles can be stated as follows:

Principle of contraction coherence (α):

$$\forall x \in X \;\; \forall S,T \in \mathrm{Pot}(X) : \left[S \subseteq T \wedge x \in S \wedge x \in C(T) \rightarrow x \in C(S)\right]$$

Principle of conjunction coherence (β):

$$\forall x \in X \;\; \forall S,T \in \mathrm{Pot}(X) : \left[x \in C(S) \wedge x \in C(T) \rightarrow x \in C(S \cup T)\right]$$

However, the principle of expansion coherence (γ) is not violated in this example:

$$\forall x, y \in X \;\; \forall S, T \in \mathrm{Pot}(X) : \left[S \subseteq T \wedge x, y \in C(S) \rightarrow \left[x \in C(T) \rightarrow y \in C(T) \right] \right]$$

We said that being polite should not be necessarily irrational. If being polite results in accepting the rule that one should not choose the most comfortable chair, following this rule should be compatible with being rational. But \¥" would also not be justified to postulate that preferences which are the expression of politeness, i.e., which you have because you wish to conform with this rule (not choosing the only most comfortable chair available) should result in a respective modification of how to rank overall states of affairs in order to save the conceptual framework of rational choice. It is the acceptance of this respective rule of politeness and not some modification of how to value states of affairs (i.e., consequences of one's actions) which motivates these preferences. It is obvious that the supposed rationality of these preferences is a result of some inadequate description. We should neither accept that preferences motivated by some rule of politeness have to be irrational nor that it can be rational to have preferences which violate the principles of expansion or contraction coherence. The way out is to redescribe these preferences motivated by this rule of politeness such that they can be rational and coherent.

This redescription can be done in the following manner: a comprehensive description of the situation considers the difference between

(1) a choice between alternatives where one of these alternatives results in violating the rule of politeness and the other doesn't and
(2) a choice between alternatives which do not include a violation of this principle of politeness.

In some cases expanding or contracting the set of alternatives can result in a change from the first type of choice to the second. In this case it would be inadequate to apply the principles of expansion and contraction coherence because the elements of the respective sets change if the set of alternatives expands or contracts. It would be inadequate because it makes a difference whether

(1) to decide between x and y in case choosing x violates a rule of politeness and choosing y doesn't or
(2) to decide between x and y in case neither choosing x nor choosing y violates a rule of politeness.

Expanding the set of alternatives may change the alternatives if they are recognized in a sufficiently comprehensive manner. A conceptual framework of rational choice which is committed to agency-consequentialism does not allow for a sufficiently comprehensive description. It precludes an adequate description of many choice-situations. In some cases it renders preferences irrational which are not irrational. To be polite is not necessarily irrational, but in some cases being polite is not

compatible with agency-consequentialism. If the alternatives are comprehensively described being polite does not violate the rationality principles α or γ.

Ramsey Coherence

Let us come back to the compatibility assumption. Is there any reason to suppose that polite preferences are not Ramsey-coherent? In fact, properties of choice functions and properties of corresponding preference relations are closely related.[37] For example: if the choice function is well-defined, contraction coherence is sufficient for rendering the corresponding preference relation acyclical. On the other hand, if the preference relation is acyclical, then the corresponding choice function is well-defined. If the conditions α, β and γ are satisfied for some well-defined choice function, the corresponding preference relation is an ordering. Therefore, Ramsey-coherence would be secured if we had no reason to assume that polite preferences violate the rationality conditions α, β and γ. To be sure, this argument does not prove that polite preferences of this kind in general satisfy α, β and γ, but since we rejected the claim that polite preferences are incompatible with basic principles of coherence there seems to be no good reason any more for doubting the compatibility assumption (given its initial plausibility).

At this stage we should include Broome's Argument and draw the line separating his and Vallentyne's view from mine.[38] Broome, too, rejects arguments opposing the view that Ramsey-coherence is a requirement for rational preferences. However, Broome (and Vallentyne) understand their rejections of these arguments as a general justification of a teleological theory of rational choice. It is not quite clear whether this kind of teleology covers what I called agency-consequentialism. If it did it would be a case of a *non-sequitur.* On the other hand, if Ramsey-coherence was an argument in favour of agency-consequentialism (the 'teleological theory'), my argument above would be proven to be wrong. I have tried to show that it is not.

Let us pick out Broome's somewhat famous example. In order to develop the general argument of his book, Broome rejects an apparent counter-example for the transitivity of preference relations. The counter-example runs as follows: Maurice prefers visiting Rome (R) to mountaineering in the Alps (M), and he prefers staying at home (H) to visiting Rome. However, he does not prefer staying at home to mountaineering; if he had a choice between these two alternatives, he would go for mountaineering. So Maurice's preferences are intransitive.[39] Maurice's claim to rationality is that mountaineering frightens him, so he prefers visiting Rome. Sightseeing bores him, so he prefers staying at home. But to stay at home when he could have gone mountaineering would, he believes, be a coward thing to do. This is the reason why he chooses mountaineering instead of the alternative to stay at home. Whereas if he chose to visit Rome instead of mountaineering this could not be taken as an expression of cowardice.

It seems that this is an example for the fact that options can change in the context of different alternatives. Staying at home is a different option depending on whether the alternative is visiting Rome or mountaineering. Let us call this phenomenon

choice-dependence of alternatives. A very general way of dealing with choice-dependence of alternatives would be to index each alternative with the given choice just as a matter of care. A closer look might then reveal that many of these indexes can be dismissed. But in the marginal case of a general indexing of alternatives the comparability of alternatives would collapse completely. If it is the choice itself which makes the difference, this tragedy would not save Ramsey-coherence (i.e., the compatibility assumption) but on the contrary would undermine its first constitutive principle and this is the connectivity axiom. We should not save transitivity by giving up another equally essential principle of preference coherence.

Here is a quick exposition of how indexing choice dependent alternatives is to be understood:

$$H > R \wedge R > M \wedge M > H \text{ is an intransitive preference relation.}$$

If H in the first choice situation is different from H in the third we can index H as 'H_R' and 'H_M' where H_R is 'staying at home if the alternative is to visit Rome' and H_M is 'staying at home if the alternative is to go mountaineering in the Alps'.

This allows for a new transitive preference relation $H_R > R \wedge R > M \wedge M > H_M$. The underlying step is a transition from the set of three alternatives {H, R, M} to the set of the four alternatives $\{H_R, R, M, H_M\}$.

One interpretation of indexed alternatives would result in a radical contextualisation: by definition, there would be *no* preference between H_M and H_R possible, because, e.g., the index M makes sense only if the preference is between staying at home and mountaineering.

Broome's example can be reinterpreted in a consequentialist fashion: the state of affairs 'Maurice is staying at home when the alternative is to visit Rome' could be different from the state of affairs 'Maurice is staying at home when the alternative is to go mountaineering in the Alps'. Maurice could, for example, feel better in the first in comparison to the second and in this way he could have the preference for H_R compared to H_M. But in fact it is not at all sure that such a consequentialist solution exists. If the situation is interpreted consequentialistically, the choice situation is only indirectly relevant for determining the respective alternatives. If mental-state-optimization was the metaphysical background of our theory of rational choice we would have to assume from the outset that a choice-dependent preference relation can be rational only if there is a consequentialist (and more specifically mental-state-) reinterpretation which would render the apparent choice-dependence as merely an indirect effect of mental-state-optimization. But we have already seen that it is not reasonable to take this stance.

Weighing Reasons

Radical contextualization undermines practical rationality. There must be something in common which gives a reason to have this preference now and that preference then. We refer to such common traits in giving reasons for action. If

choice-sensitivity is rational, there must be some reason for being choice-dependent. One reason could indeed be that the agent wants to optimize her mental state. This reason does not only render choice-dependence indirect (or not geniune) but also allows for a wide comparability of alternatives: if the agent expects to feel better if the first of two alternatives is realized she will opt for the first, independently of the choice situation. If the choice situation has an influence on how the agent feels, this effect has to be included into a full description.

If a reason is responsible which is directly related to the choice situation as in the example above (choosing chairs), it offers another track of comparability beyond the context of the situation of choice. To act on the basis of reasons secures comparability. Weighing reasons generally secures universal comparability. The coherence requirement of completeness is prima facily plausible not only within the traditional consequentialist perspective of weighing goods (i.e., values of outcomes) but also in the broader perspective of weighing reasons.

The account of rationality which is to be defended here can be understood as a proposal of *weighing reasons* – instead of *weighing goods.* Reasons are the fundamental elements from which a theory of practical rationality has to be constructed. In order to choose coherently we have to weigh reasons and this includes considerations of probability. The account of weighing reasons differs from traditional deontological acounts insofar as it postulates preference coherence (i.e., quantitative representability of rational epistemic and conative attitudes by real-valued belief- and value-functions) and it differs from consequentialist (or teleological) accounts in so far as it takes reasons into account which can not be reduced to optimizing state of affairs or outcomes.

Most of our good prima facie reasons for action entail genuine choice-dependence. The outcomes alone, i.e., the resulting states of affairs do not determine whether the choice conforms with the respective reason. Therefore, the comparability of alternatives can only be saved if they are comprehensively described, i.e., described such that their traits which are relevant for judging whether the choice conforms to the reason are included. If alternatives of this kind are described adequatly the reason in question is part of the description. If we compare alternatives and the reason in question is part of their description we implicitly have to weigh reasons.

Structural Rationality

I conclude with a programmatic outlook on an account which I call ‘structural rationality’. It integrates non-consequential reasons for action and does not touch the core of rational choice theory.[40] It starts from the assumption that point-wise outcome-optimization is (a) incompatible with many of our good reasons for action, (b) would desintegrate the structure of interaction on which social cooperation is based and (c) would desintegrate the person. Act-utilitarians argue that this assumption would not be true for perfectly rational utilitarian agents: the ideal utilitarian agents – they think – could adopt agency-consequentialism without destroying the structure of interaction on which social cooperation is based and without

desintegrating their own personality. For act-utilitarianism the conflict with seemingly good reasons for action is merely an indication for the fact that our moral intuitions are not reliable. For the purpose of my argument it is not necessary to prove the truth of this claim.[41] For even if the ideal moral person in the sense of act-utilitarianism could live on the basis of point-wise maximization, it would still be true that the rest of humankind could not. The rest of humankind differs in ranking states of affairs and weighing reasons, in personal projects and interests. A general theory of practical rationality should not presuppose that the overall value functions representing revealed preferences are interpersonally identical.

A general theory of practical rationality should not be too restrictive. It should not prescribe which personal projects we try to realize in our lifes, with whom we want to cooperate, which rules we want to follow, which reasons are the most relevant for our form of life. Some element of Humeanism is healthy for a general theory of rationality: what we are aiming at, how our conative attitudes are constituted is largely up to the acting person. If a person wants to be polite and therefore accepts a certain rule of politeness we should not call this person irrational on grounds of the respective rule not being outcome-oriented. We saw above how it is possible to integrate such a conative attitude into a general framework of rational choice. Radical Humeans would stop here and refuse to go further, because they think that there is no way to develop rational criteria for ends.

The theory of structural rationality starts where radical Humeans stop. The idea is that – in analogy to descriptive beliefs which are neither isolated elements of an unstructured set – our actions reveal a vast system of subjectively accepted reasons which are closely interconnected and integrated. In order to conform with such a structured system of reasons we have to be careful in choosing our actions such that they conform with it as a whole. Point-wise outcome-optimization would result in a form of life which would not accord with the conative attitudes of most persons. It is perfectly rational to refrain from point-wise optimization because you do not wish to live the life which would result.

It is helpful to introduce the idea of structural rationality in a strictly consequentialist framework (i.e., the only relevant values would be related to outcomes or state of affairs which are caused by actions) emphasizing the tension between point-wise-optimization of consequences and overall optimization of consequences. The modification then would be to give up agency-consequentialism in order to save overall-consequentialism. But this first step of the analysis opens the door for a more comprehensive understanding of practical rationality which departs from consequentialism.

The crucial point of the idea of stuctural rationality is the relation between singular actions and overall preferred structures of agency. If you prefer a form of society in which politeness plays a role and if a specific rule is constitutive for your understanding of politeness, then this has an influence on your rational choice behaviour. Even if your personal rule-conformity or rule-defection has no influence on the respective social states and processes as a whole, it is not necessarily irrational to act in conformity with the preferred structure of social action. The conative

attitude referring to a social structure has to be in coherence with the respective conative attitude in the singular case.

Acts are individuated modulo intentions, and intentions are individuated modulo reasons. In many cases, an agent's intention does not refer to a singular action. A whole sequence of actions can be necessary to act on the intention. This sequence as a whole can again be described as an action modulo this 'embracing' intention. Depending on the agent's intentions, some actions are performed by performing other, more fine-grained actions. The minimal units are constituted by the intentional state of the agent in the moment the action is done. There is no agent-relative or even universal repertoire of 'basic actions'. Which actions are basic depends completely on the respective choice situation and the intentional state of the agent in this situation. The more experienced the agent is the bigger are her basic action units. The basic actions are chosen within the broader framework of the agent's embracing intentions. The intention to do x_1 is motivated by the intention to do y because you have the intention to do y and you know that doing x_1, x_2, X_3, ..., x_N would result in doing y.

The relation between fine-grained and embracing intentions should be understood without falling back into a consequentialist interpretation. The relation is not one of causal determination of outcomes. Structural traits of rationality are constitutive parts of the intentional sets guiding our actions. An adequate theory of practical rationality is not possible without taking them into consideration. As shown in this paper, this cannot be done within the conventional consequentialist interpretation of rational choice. But if this biased interpretation is given up, structural traits are compatible with Ramsey-coherence, the core element of rational choice theory.

References

Austin, J.L. 1955. *How to Do Things With Words.* Oxford: Clarendon.

Broome, J. 1991. *Weighing Goods. Equality, Uncertainty, and Time.* Oxford: Basil Blackwell.

Gowans, C.W. (ed.). 1987. *Moral Dilemmas.* New York/Oxford: Oxford University Press.

Kern, L., and J. Nida-Rümelin. 1994. *Logik kollektiver Entscheidungen.* München: Oldenbourg.

McClennen, E.F. 1990. *Rationality and Dynamic Choice.* Cambridge: Cambridge University Press.

Nida-Rümelin, J. 1991. Practical Reason or Metapreferences? An Undogmatic Defense of Kantian Morality. *Theory and Decision* 30: 133–162.

Nida-Rümelin, J. 1993. *Kritik des Konsequentialismus.* München: Oldenbourg (paperback ed. 1995).

Nida-Rümelin, J. 1997. Why Consequentialism Fails. In *Contemporary Action Theory*, Vol. II, ed. G. Hilström-Hintikka and R. Tuomela, 295–308. Dordrecht: Kluwer.

Nida-Rümelin, J. 1998. Subjective and Objective Reasons. In *The Role of Pragmatics in Contemporary Philosophy*, ed. P. Weingartner, G. Schurz and G. Dorn, 368–379. Wien: Hölder-Pichler-Tempsky.
Searle, J. 1969. *Speech Acts.* London: Cambridge University Press.
Sen, A.K. 1971. *Collective Choice and Social Welfare.* San Francisco: Holden Day.
Sen, A.K. 1977. Social Choice Theory: A Re-examination. *Econometrica* 45: 53–89.
Sen, A.K. 1997. Maximization and the Act of Choice. *Econometrica* 65: 745–779.
Vallentyne, P. 1988. Gimmicky Representations of Moral Theories. *Metaphilosophy* 3&4: 253–263.

Structural Rationality and Collective Intentions

Structural Rationality

The central assumption of structural rationality is that *an act can be rational if it is based on the motivation to conform to a certain pattern (or "structure") of actions.* Note that I do not claim that the pattern must be optimal in order to render the act rational. From the agent's perspective, the behavioral pattern need not be the best of all possible patterns. We will see below why good reasons count in favor of this weaker claim. But before we look at this notion in more detail, let me give a simple (intrapersonal) example of what I understand by "structurally rational."[42]

Think of an artist – call her A – who is convinced that she should reserve certain hours of her day for her work, say, from 9:00 in the morning until 7:00 in the evening, with a break between 1:00 and 3:00 in the afternoon. Every day she sits in her studio working, and she continues to do so her whole life. I want to focus on the kind of motivation in play here, that is, to get up in the morning and begin to work day by day without having any external pressure. Sometimes A has simply no inclination to start working early in the morning. In other words, she then does not fulfill a present wish like "I desire to work" when she goes to work.

If one acts, one always has a desire on which one acts. The desire on which the artist acts is the desire to conform to a temporal structure. This is not a wish in the sense of a Kantian inclination ("Neigung"). Instead, the artist is motivated by a concrete temporal structure she has imposed on her work. Motivations such as this fall under the concept of *structural rationality;* on the assumption of structural rationality, they are not necessarily irrational. To choose an action because it conforms to a structure of actions (structural intentions) can be rational.

The concept of *revealed preference* states that preferences are shown as what they are whenever they are revealed by actual choice. In the example given above, the choice is to act in conformity with a temporal structure. The choice is not made because the action optimizes consequences. To interpret it thus would be incompatible with how this choice is motivated. If somebody postulates that an act can be rational only if its motive is to optimize consequences, acting in conformity with a temporal structure would be irrational.

One might say that it is not tfie *motivation* to optimize consequences that is decisive for its rationality but the *fact* that an action optimizes consequences. Therefore, the artist's decision to begin her work every day at 9:00 in the morning may be rational although its motive is not to optimize consequences. According to this view, a decision may be rational if, regardless of its structural motivation, it in fact optimizes consequences. For example, one could assume that a single defection would threaten the whole *form of life,* for it might cause further defections, which result in reduced artistic productivity. But mere conformity to the rule (the temporal struc-

Published in S.R. Chant, F. Hindriks and G. Preyer, *From Inidvidual to Collective Intentionality – New Essays*, Oxford University Press, Oxford, pp. 207–222.

ture of work) cannot make an action rational. But what if the artist answers: "It is exactly my desire to act in conformity with this specific temporal structure"? We should accept the individual's avowed or manifest preferences.

An act is rational if there are good reasons for performing it. Structural rationality takes this into account. *It does not dismiss the plurality of good reasons;* it does not try to reduce the plurality of practical reasons to a few or even to one. The account of structural rationality is not *radical* and not *rationalist* in the sense of trying to replace lifeworld complexity by a simple, all-covering criterion of rationality.

Only a few types of lifeworld practical reasons are of the consequentialist land. If I try to do my best to reduce the suffering of an ill person, I act on the basis of consequentialist reasons – in this example, altruistic consequentialist reasons. The consequence I am trying to optimize is the minimization of the time of an ill persons suffering, and it might well be possible to reconstruct this motivation by means of a quantitative utility function or, if the consequences of my actions are uncertain, a utility function combined with a quantitative probability function. Even this consequentialist motivation, however, follows (mostly implicit) rules and is *embedded within patterned traits of social interaction.* The altruist agent of our example might, for example, accept rules like "do not change the essential psychological traits of the suffering person by medication."

The vast majority of reasons for action are not directed toward improving a state of affairs, whatever the measure of quality of a state of affairs might be. I give three examples of nonconsequentialist lifeworld reasons:

(1) We have a good prima facie reason to perform an action if we promised to do it. This reason is not directed toward improving or even optimizing a state of affairs.
(2) That somebody asks me to do something constitutes a good reason to do it. What constitutes the reason for the action is not that it improves a state of affairs but rather that I am asked to perform it. This is true even if the person asking was motivated by improving some state of affairs.
(3) As a parent, I have duties regarding my children. These duties involve many types of actions, but only a few of them can be understood as optimizing some state of affairs.

There are duties that are connected with social roles, obligations that are related to personal commitments, and still others that originate from prescriptions or rights. The idea that all of these prima facie duties could be reduced to the normative criterion of optimizing states of affairs has a rationalistic taste of the worst land: dismiss all experiences, forget all lifeworld practices, and begin totally anew. Postulate some principles and try to deduce the rest of your theory. Compared to this, the account of structural rationality is modest. It avoids philosophical and theoretical hypocrisy; it takes practices and persons seriously but, nevertheless, preserves the mathematical core of decision theory.[43]

In the intrapersonal case, we act rationally if the series of singular acts constitutes a structure (patterns of actions) that is preferred by the agent or, in a weaker

formulation, if it forms a pattern of acts that the agent can accept as part of her *form of life*. Philosophical or economic theory should not prescribe which kind of practical reasons constitute these structures as elements of a chosen form of life. If a person only acts on her present inclinations, she will (in the long run) not be happy with the result. The form of life resulting from following your present inclinations, that is, always trying to fulfill your wishes the moment you have them, will not be a good one. A flourishing life requires structurally rational action, that is, a practice that results in a well-structured life. Education and, to put it more broadly, culturalization are about how to structure ones life and how to act in a structurally rational way. Philosophy or economic theory should not collide with these lifeworld experiences.

The *inter*personal case of structural rationality is closely related to the *intra*personal case. Take *one* person acting at *different* points of time like *different* persons acting at *one* point of time. Let me explain this with the help of an intrapersonal example. A traveler is about to visit Rome in two weeks. She is taking Italian lessons because she expects she'll be better off in Rome if she can communicate in the local tongue. She does not have the preference to communicate now, but only an expectation that communication will become relevant in the future. At the moment, she has no inclination to attend the Italian lessons.[44] In Rome then, she will have the desire to communicate and, consequently, would do a lot of things to fulfill it. Now, however, she only has the expectation that she will have that desire in Rome and prefers to hang out in a coffee shop drinking a latte instead of attending the Italian lesson. If you take only actual desires (inclinations) into account, utility functions would change a lot over time.

A paradigmatic case of structurally rational agency is cooperation: I perform an action because it conforms to an interpersonal pattern (n-tuple) of actions (s) that I can approve of. If the cooperation succeeds, every participant contributes his part, expecting that all others do their part, in order to realize the collective intention to collectively do (s). This is compatible with the existence of other patterns (s') (collective actions in the formal sense of an n-tuple of individual acts) that I would prefer to (s). Therefore, I cooperate in doing my part if I can expect the others to do their part, knowing that if everybody does what is individually optimizing, all fare worse, and also knowing that if everybody does what could be part of the pattern that is best for himself, cooperation would not succeed. In successful cases, cooperation is based on mutual expectations that can be conceptualized as common knowledge within the respective group. Hence, I do my part because I expect others to do their part, and if the others have the respective expectations, cooperation takes place; the jointly intended collective practice is realized successfully.

This introduces a *normative dimension* into collective action. I cannot expect that others do their part in a collective practice that they take to be unfair. There must be a basic normative consensus that one can expect others to do their part. Expectation is not just an expression of subjective probability, but also an expression of implicit normative attitudes. I can expect you to do your part in a common practice only if this common practice is acceptable for both of us.

Cooperation

I take cooperation to be a specific form of interaction that is defined by the prisoner's dilemma game format. Cooperation is different from mere coordination. For example, if two persons have a common interest to meet but are indifferent about where to meet, they usually communicate to fix where and when. In game theory, such a situation is called coordination game. There are coordination games with and without conflict between the participants A and B (cf. Tables 1 and 2). The coordination game with conflict is defined by interpersonally differing outcomes of various solutions such that at least one of the participants prefers a different solution than the other participant.

In contrast to coordination games, there is a singular collective action composed by individually optimizing strategies in cooperation games. Even more, for every individual, there is one dominant strategy, for example, a strategy whose individual results are better than the results of any other strategy, independent of what the other participants do. The problem is that individual optimization results in an equilibrium that is not Pareto-efficient.[45] At least one other combination of individual actions exists that is better for all participants (see Table 3).

Table 1 Coordination game without conflict

		B	
		b_1	b_2
A	a_1	1 / 1	0 / 0
	a_2	0 / 0	1 / 1

Table 2 Coordination game with conflict

		B	
		b_1	b_2
A	a_1	1 / 2	0 / 0
	a_2	0 / 0	2 / 1

Table 3 Prisoner's dilemma (PD)

		B	
		b_1 (C)	b_2 (D)
A	a_1 (C)	3 / 3	1 / 4
	a_2 (D)	4 / 1	2 / 2

Probably there is no other game format that has been discussed so intensively in the last five decades as the prisoner's dilemma. Nevertheless, it remains a challenge. Many solutions to it have been proposed. The most prominent of them bears on the standard rc-paradigm according to which there is only one rational strategy in the prisoner's dilemma – namely, defection (D) (cf. Tables 4 and 6). The fact that the combination of defective acts is Pareto-inferior should then motivate the participants to establish institutions that, via sanctions, change the outcome of the game in a way that cooperation (C) becomes optimizing (c.f. Tables 5 and 7).

Table 4

		B	
		b_1 (C)	b_2 (D)
A	a_1 (C)	3 / 3	1 / 4 (-2)
	a_2 (D)	4 (-2) / 1	2 / 2

Table 5

		B	
		b_1 (C)	b_2 (D)
A	a_1 (C)	3 / 3	1 / 2
	a_2 (D)	2 / 1	2 / 2

Table 6

		B	
		b_1 (C)	b_2 (D)
A	a_1 (C)	3 / 3	1 / 4 (-2)
	a_2 (D)	4 (-2) / 1	2 (-2) / 2 (-2)

Table 7

		B	
		b_1 (C)	b_2 (D)
A	a_1 (C)	3 / 3	1 / 2
	a_2 (D)	2 / 1	0 / 0

If B cooperates and A defects, A's defection is punished (−2); the same holds for defection by B if A cooperates:

The result is a game format with two equilibrium points:

If defection is punished independently of what the other person does,

there remains only one equilibrium point (both cooperate):

The account of structural rationality is not at odds with *political* solutions of cooperation dilemmas establishing institutions that change outcomes via sanctions so that defection is not optimizing anymore (strictly speaking, *defection* (D) ceases to exist as a strategy if the individual payoffs are changed). But such institutional solutions to the cooperation dilemma are rare. The prisoner's dilemma is endemic in human interaction because *rules of any kind usually result in an n-person prisoners dilemma.* This is true even for the constitutive rules of language like *truthfulness, trust,* and *liability.*[46] Rules that go beyond mere coordination require a higher degree of conformity in order to establish stable patterns of interaction. Individual optimization of outcomes (i.e., of a state of affairs) is incompatible with this high degree of conformity. Rule following – be it explicit or implicit – has a normative motivation: I do my part because I approve of the patterned practices if I can expect that all or most of the group members conform to them. The most interesting kind of rule following is based on the empirical and normative phenomenon of structural rationality. Cooperation, I claim, is a paradigmatic case of structural rationality, because cooperation exhibits the basic elements of the concept of *structural rationality.* Structural rationality postulates that one should not try *to translate structural intentions into modifications of utility functions over outcomes within* a decision-theoretic framework.[47] Rather, the concept of structural rationality allows for a new description of the dilemma of cooperation. It takes the motivation of cooperating persons seriously, that is, it accepts the motive to act as a good (cooperative) reason. A cooperative reason of this kind can be described as follows: a person has a good cooperative reason to contribute to a specific collective action, even though, independent of the contribution of other persons, it would be better for her as an individual not to do so. Reasons of this kind are, *per definitionem,* cooperative reasons. My account thus simply drops the consequentialist dogma that rational action is directed toward optimizing states of affairs (i.e., in the standard rc-paradigm, optimizing the *subjective, individual* state of the agent).[48]

Seen from the perspective of structural rationality, contributing to a collective action that one approves of can be a good reason for action. Structural reasons of this kind motivate persons many times a day. It does not make much sense telling them that this motivation is irrational or that the alleged reason is not a genuine one. However, if we accept *cooperative reasons,* we do indeed run into game-theoretic troubles, because they speak in favor of choosing strategies (within the prisoners dilemma format) that the individual does not prefer. Note that the format of the prisoner's dilemma is defined by the individual rankings of collective actions (i.e., strategy combinations).

Person A: DC > CC > DD > CD
Person B: CD > CC > DD > DC

Philosophical friends of cooperation who are also friends of standard rational choice theory try to combine both of these ideas by introducing utility transfers, which model cooperative behavior as altruistic motivation or a common ethos (i.e., a common system of rules and norms as well as a specific feeling of belonging).[49] And indeed, it seems to be an attractive idea to solve the problem of cooperation by embedding it within the broader frame of altruistic motivation. But in doing so, we are on the wrong track, although every prisoner's dilemma can be resolved by assuming the necessary transfer rate of utilities to overcome its format. In fact, there is only one way to universally avoid a conflict between individual optimization and cooperation, and that is to give all utilities involved equal weight. Let us call it the "utilitarian solution."[50] However, if we give our own utilities more weight than those of others, we cannot overcome all PD-situations. They will remain PD-situations even after the transfer. If, conversely, we assign more weight to the utilities of others than to our own, some PD-situations will be transformed into a "Saint's dilemma": The Saint's dilemma is the reverse format of the prisoners dilemma. Being overtly altruistic, the saint ends up in a Pareto-inferior collective action if they individually optimize outcomes. Utilitarians might find this result attractive since it seems to show that there is only one way of overcoming the cooperation dilemma, and this is to act on the basis of utilitarian principles.

However, I think that this result can be interpreted in both ways, speaking either in favor of utilitarianism or in favor of a plausible (structural) account of cooperation. In order to be able to be cooperative, individuals do not need to develop a strong altruistic and egalitarian motivation (i.e., giving equal weight to every person's utilities), because they are able to cooperate directly, without changing their preferences. The utilitarian solution avoids genuine cooperation, but there is no cultural or even genetic mechanism that forces human individuals *not* to cooperate. On the contrary, there may be genetic and cultural barriers against a utilitarian motivation.[51] Collective action in general amounts to more than the results out of individual optimizing strategies.

It is not possible to integrate cooperative rationality easily within the conceptual frame of standard decision and game theory. Optimizing outcomes is the criterion of rationality in decision and in game theory, even if there are considerable differences between these two branches of rational choice.[52] Applied to individual actions and strategies, the predicate *rational* is expressed by a function with two decision-theoretic variables, namely, subjective probability and subjective utility. In standard game theory (in contrast to the Bayesian variant of game theory), optimization becomes a function of the individual's utilities, that is, outcomes alone. However, in addition, a meta-criterion requires that every player (i.e., every participant in an interaction) has the same knowledge regarding the game format and the rationality criterion, that is, the solutions of the game. This supplement to decision theory singles out the Nash equilibrium as the solution of standard game theory.[53] If you translate the game-theoretic format of the Prisoner's dilemma into a decision-theoretic one, there can be no doubt that a rational person chooses D, because D is the dominant strategy (i.e., the strategy that has better results whatever the circumstances). Translated into a decision-theoretic situation, i.e., a situation in which the actor is confronted not with other persons' strategies but with natural events, domi-

nant strategies are the only rational ones (provided they exist). From this observation, one may deduce that choosing D in a PD-situation amounts to treating the other player like a natural event. One might say that if there is no causal relation between the players' strategies, one should indeed treat the other players like natural events. Think of an analogy to this decision-theoretic translation. If, for some reason, the probabilities of the circumstances depend on your choice of strategies, the dominant strategy might not be rational.

The essential point of my argument is that cooperative action is motivated by specific types of reasons. Let us call them *cooperative reasons.* Like most reasons, they are constituted by an interpersonal relation. I have a cooperative reason for acting if I want that we, the co-operators, contribute each individually to some collective action that we jointly prefer. This *jointly* is to be understood collectively, not distributively. This does not deny that the co-operators may not know any better collective action. It merely means that we – as the group of co-operators – have the common normative belief that this collective action is acceptable to all of us.

To get a better understanding of this form of collective acceptance, it is useful to look at *collective choice theory.* Its basic concept is an aggregation function that attributes collective preference relations to a collection of individual utility functions. Aggregations of functions attribute to each collection of individual preference orderings[54] that form collective preference orderings. They are called *social welfare functions*;[55]

Cooperation takes place if there is (1) a cooperative attitude and (2) a commonly approved collective action and (3) if the strategy combinations are such that there is an n-person prisoner's dilemma (possibly with several cooperative strategies, i.e., strategy combinations that are Pareto-superior to the result of individual outcome optimization). All rational forms of preference aggregation are Pareto-inclusive; that is to say, if a collective action is Pareto-superior to another, it is collectively preferred. Pareto-inclusiveness is compatible with many different normative accounts of aggregation. For example, the famous difference principle developed by Rawls in his *Theory of Justice* (1971) is Pareto-inclusive in this way. Even the principle of majority vote is Pareto-inclusive as is the unanimity-principle; even more, the latter allows collective decisions only in case of Pareto-superiority (note that it is not the case that any Pareto-optimal outcome results in unanimity) (Table 8).

a_1/b_1 and a_2/b_2 are each Pareto-optimal collective decisions. However, there is no consensus between A and B about which of these two Pareto-optimal collective decisions should be realized. A prefers a_1/b_1 and B prefers a_2/b_2. Pareto-optimality therefore does not result in unanimity.

Table 8

		B	
		b_1	B_2
A	a_1	2 / 1	0 / 0
	a_2	0 / 0	1 / 2

Hence, I conclude that cooperation is enhanced by any generally accepted reasonable rule of aggregation because of its Pareto-inclusiveness. To put it differently: it is a normative consensus shared by all individuals that only cooperative collective actions can be jointly approved of; that is, only cooperation is collectively rational.

The impossibility results of collective choice theory (the theorems of Arrow, Gibbard, Satterthwaite, and Sen) and the "breakdowns of rationality" in game theory (Howard Nigel) suggest that there is no universal rule to identify collective rationality. More specifically, there is no universal rule to identify the cooperative actions that can be jointly approved of. It is up to institutional arrangements, social relations, cultural traditions, and moral judgements to choose a cooperative action that can be jointly approved of. Equity-, fairness-, and efficiency-conditions play an important role in selecting these cooperative collective actions.[56]

The reorientation of game theory, as suggested by Thomas C. Schelling,[57] proved to be quite successful in its application to international relations. It can also be interpreted as a theory of reducing the plurality of possible cooperative collective actions by introducing the criterion of *salience.* At first sight, salience is not a normative concept. However, in bargaining processes, salience can be used as a remedy to individual optimization and the instabilities and indeterminacies resulting from it. We should agree on a salient solution because many possible cooperative solutions exist, and none of them are more justified than the other. It is easier to reach a consensus if some cooperative solutions seem somehow outstanding. The normative "ought" of cooperation is thus transferred to the salient solution. We say: "If we both want to cooperate, we should prove it by both consenting to the salient solution." Egalitarian cooperative collective actions fulfill both criteria, that is, that of equity and that of salience. For example, the division of the cake among five hungry children amounts to an infinite set of Pareto-optimal alternatives. Every division of a cake in five parts is Pareto-optimal if all of the children are sufficiently hungry. The equal division is salient even if the children differ in weight, appetite, or age. It is possible to model the situation as a prisoner's dilemma. Every child may hope that if he screams louder than the others, he will get a larger part of the cake; some of the children may think that they have priority because of some reason or other. But, probably, they all know that there is only one general consensus possible, which is sharing the cake equally. The D-strategy refers to the conditions in which every child demands more than his equal share, and the C-strategy refers to the condition of behaving civilized, and thus being content with one's equal share.

Cooperation is also often made possible by a consensus on second order rules. The participants each have their own first-order rules of how to evaluate feasible collective actions. Some may place special weight on equity, others on efficiency; and some may combine these two principles in a reasonable way. In many cases, it would be impossible to reach a consensus on the basis of first-order rules because divergent opinions exist on exactly how to evaluate and judge different alternatives. However, the participants may agree on second-order rules, which help to find a cooperative collective action that is acceptable for everyone. For instance, simple majority vote is one of these second-order rules. If we cannot reach a consensus about which of the different feasible cooperative collective actions should be performed, we vote and everybody performs her action as a result of this collective action.

Collective Intentions

In the previous sections, I argued that a specific type of collective intentionality is indispensable for cooperation:

(1) Every (potential) participant must have a cooperative attitude – that is, she must be willing to fulfill her part in an act of performing a cooperative collective action.
(2) There must be common knowledge regarding (1), that is, every group member must know that every other member intends to fulfill her part of the collective action.
(3) There must exist at least one cooperative collective action (i.e., a combination of individual strategies that has a better result for everyone involved than the result of individual optimization) that is acceptable for all (potential) co-operators – hence, (3) is a common normative belief.
(4) There must be common knowledge regarding (3).
(5) There must be a consensus that yields one of the feasible cooperative collective actions as acceptable for all. However, as shown in section 9.2, no abstract universal criterion exists to solve the uniqueness problem.[58]

Individual cooperative action requires a certain detachment from the personal point of view; it requires that I treat my potential co-operators as different from natural events. Cooperative reasons address all potential co-operators. There are no solipsistic reasons. My desires, attitudes, and interests do play a role, but desires are always tied to attitudes and interests of others. In collective actions, the actions of other group members – as well as my own actions – are understood as part of a common practice. To act cooperatively would not make sense (and a fortiori would not be rational) if it did not include an intention to contribute to a collective action. This intention is constituted by common normative and empirical knowledge; it includes mutual expectations that are interpersonally coherent.

The kind of collective intentionality that is constitutive of cooperation does not necessarily require a community or a group that is bound together by a specific ethos. Think of a society in which a strong individualism prevents the development of any sense of cultural, social, or ethnic community. If the members of this society are asked which community they belong to, they would answer "none." If communities are understood as being held together by a common ethos, that is, a common system of rules and norms and a specific feeling of belonging, we still can imagine a society in which there are no communities of this kind but in which cooperation among individuals nevertheless exists. The basis of this cooperation may be trust, interest, and some common normative beliefs regarding autonomy and fairness. For the salce of the argument, let us suppose that all individuals share these same principles of individual autonomy and fairness. This will enable them to cooperate on the basis of a general cooperative attitude (which we have shown can be rational) as well as a common knowledge regarding the persons and interests involved. Also, the selection of a unique cooperative collective action is included, which fulfills the criteria of individual autonomy and fairness and which is selected – if necessary – by a second-order rule like majority vote.

Note that this interpretation is only possible on the basis of my account of structural rationality. The reason for this is that cooperation without community, that is, without group ethos, does not introduce any changes in values or utilities. The participants have only one type of evaluation, which is represented by the game format of the interaction. The only additional assumption one needs is that the persons do not optimize outcomes but accept a specific type of interpersonal structural intentions (i.e., cooperative intentions). These intentions can be described by referring to the respective game format. They cannot be described by referring to the outcomes of each participant's strategic options. Thus, we drive a wedge between choice and preference because strategy combinations cannot be identified with outcomes anymore. Outcomes retain their original status as representing subjective consequences.[59] Preference orderings over outcomes are not mirrored by preference orderings over the respective strategy combinations. In the prisoner's dilemma case, this means that A may prefer CC to DC, that is, A may act cooperatively if A expects B to act cooperatively, although A prefers <4,1> to <3,3>. The latter preference may be revealed if we ask A if he prefers (1), that is, receiving four and B receiving one, or if he prefers receiving three and so does B. The account of structural rationality allows A to answer (1), although he acts cooperatively if he expects B to act cooperatively, that is, he chooses C if he expects B to choose C. Since I define structural reasons as guiding rational actions, I do not need to wait for utility changes in order to render cooperative action rational. In other words, we dismiss the narrow rationality criterion of outcome optimization in order to be able to integrate structural reasons, structural intentions, cooperative action, and collective intentionality into the decision-theoretic conceptual frame.

In the previous sections, I have described cooperative action as an exemplary case of collective intentionality without any reference to group ethos or community values.[60] In this case, there is genuine collective intentionality if individual optimization of outcomes would lead to a different result than individual rational behavior. Collective intentionality is constituted by common knowledge – a common cooperative attitude and the selection of an acceptable cooperative collective strategy as described above. No transformation process is presupposed that changes values or utilities. This exemplary case can be generalized in the sense that there is no specific linkage between the cooperating partners – there is no WE and THEY; no collective identity of whatever kind is involved.[61] Cooperation comes up whenever the interaction is of the respective type and the conditions (l)-(5) of the scheme above are fulfilled. Thus, cooperation in the specific way of collective intentionality comes up whenever the circumstances allow for it. The constitution of groups, communities, and collective identities; the development of a group ethos; and the discrimination between WE and THEY are not necessary in order to render cooperation rational.

Even if we alter these conditions and allow for different forms of collective intentionality that are based, for example, on different normative judgments in different groups of agents, we can uphold the assumption that values and utilities remain unchanged. Normative judgments that are constitutive for some groups of co-operators but not for others are exclusively deontological. They refer to the rules that make cooperation possible, that is, the rules that pick out the acceptable cooperative collective action from the set of feasible cooperative collective actions but do not result in changing the evaluation of outcomes. In a way, the agents remain as they are. If subjective probabilities and utilities characterize an agent so long as his beliefs and values are sufficiently coherent, then, by cooperating, structural aspects of rationality are added without changing the agent.

References

Arrow, K. 1963. *Social Choice and Individual Values,* 2nd ed. New York: Wiley.
Davidson, D. 1973. Radical Interpretation. *Dialectica* 27: 314–328.
Douglas, M. 1986. *How Institutions Think.* Syracuse: Syracuse University Press.
Fleurbaey, M., and A. Trannoy. 2003. The Impossibility of a Paretian Egalitarian. *Social Choice and Welfare* 21: 243–263.
Gilbert, M. 2000. *Sociality and Responsibility: New Essays in Plural Subject Theory.* Lanham: Rowman & Littlefield.
Lewis, D. 1969. *Convention: A Philosophical Study.* Cambridge: Harvard University Press.
———. 1974. Radical Interpretation. *Synthese* 23: 331–344.
Margolis, H. 1982. *Selfishness, Altruism, and Rationality: A Theory of Social Choice.* Chicago: University of Chicago Press
Nagel, T. 1970. *The Possibility of Altruism.* Princeton: Princeton University Press.
Nida-Rümelin, J. 1993. *Kritik des Konsequentialismus.* München: Oldenbourg.
———. 1994. Rational Choice. Extensions and Revisions. *Ratio* 7: 122–144.
———. 1997a. *Economic Rationality and Practical Reason.* Dordrecht: Kluwer.
———. 1997b. Why Consequentialism Fails. In *Contemporary Action Theory*, vol. II, ed. Ghita Holmström-Hintikka and Raimo Tuomela, 295–308. Dordrecht: Kluwer.
———. 2001. *Strukturelle Rationalität.* Stuttgart: Reclam.
———. 2000. Rationality: Coherence and Structure. In *Rationality, Rules, and Structure*, ed.J. Nida-Rümelin and W Spohn, 1–16. Dordrecht: Kluwer.
———. 2005a. *Über menschliche Freiheit.* Stuttgart: Reclam.
———. 2005b. Why Rational Deontological Action optimizes Subjective Value. *Protosociology* 21: 182–193.
———. 2011. *Verantwortung.* Stuttgart: Reclam.
Pazner, E.A., and David Schmeidler. 1978. Egalitarian Equivalent Allocations: A New Concept in Economic Equity. *Quarterly Journal of Economics* 92*:* 671–687.
Rawls, J. 1971. *Theory of Justice.* Cambridge: Harvard University Press.
Rescher, N. 1975. *Unselfishness. The Role of the Vicarious Affects in Moral Philosophy and Social Theory.* Pittsburgh: University of Pittsburgh Press.
Savage, L.J. 1954. *The Foundations of Statistics.* New York: Dover Publications.
Schemer, S. 1982. *The Rejection of Consequentialism. A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions.* Oxford: Oxford University Press.
Schelling, T.C. 1960. *The Strategy of Conflict.* Cambridge, MA: Cambridge University Press.
Searle, J. 1969. *Speech Acts: An Essay in the Philosophy of Language.* Cambridge: Cambridge University Press.
Sen, A.K. 1977. Rational Fools; A Critique of the Behavioral Foundations of Economic Theory. *Philosophy and Public Affairs* 6: 317–344.
———. 1986. Social Choice Theory. In *Handbook of Mathematical Economics*, vol. 3, ed. K.J. Arrow and M.D. Intrilligator. Amsterdam: North Holland.
———. 1997. Maximization and the Act of Choice. *Econometrica* 65: 745–779.

Shapiro, J. 2006. Genome Informatics: The Role of DANN in *Cellular Computations*. *Biological Theory* 1: 288–301.
Spohn, W. 2012. Reversing 30 Years of Discussion: Why Causal Decision Theorists Should One-Box. *Synthese* 187: 95–122.
Tuomela, R. 2007. *The Philosophy of Sociality: The Shared Point of View.* Oxford: Oxford University Press.
Varian, H. 1974. Equity, Envy, and Efficiency. *Journal of Economic Theory* 9*:* 63–91.

Notes

1. The critique of consequentialism presented in this paper is developed in much greater detail in Nida-Rümelin (1993). The formal parts are presented in Kern/Nida-Rürnelin (1994).
2. See Kant (1785, Second Part).
3. See Habermas (1981).
4. For a formal treatment of this notion see the appendix to this paper. Cf. also Nida-Rümelin (1993, § 8), and Kern/Nida-Rürnelin (1994, Ch. 2).
5. See Vallentyne (1987, 1988) and Broome (1991).
6. For this point cf. McClennen (1990).
7. For the following argument, see also Nida-Rümelin (1993, § 51).
8. Cf. Ross (1930).
9. See e.g. Harsanyi (1955).
10. Cf. e.g. Scheffler (1982), the contributions in Scheffler (1988), and Margolis (1982).
11. See the idea of the market as a morally free zone in Gauthier (1986, ch. IV).
12. For a formal statement of Sen's theorem (commonly known as the 'Liberal Paradox' — the original proof was published as Sen (1970)) see the appendix to this paper. Additional material is contained in Kern/Nida-Rümelin (1994, Ch. 11).
13. See Gibbard (1973), Satterthwaite (1975), and the appendix to this paper.
14. This is not – as it might seem – a circular definition, because the predicate 'cooperative' in the last sentence has no other function than to identify the type of action. One could substitute 'cooperative'by 'respective'.
15. The idea of analyzing metapreference structures in the context of the prisoners' dilemma was introduced by Sen (1977), see also Nida-Rümelin (1991). For a formal definition of the prisoners' dilemma see also the appendix to this paper.
16. I developed the arguments in favour of a deontological account of rationality and morality in *Kritik des Konsequentialismus,* München/Wien, second ed.1995, in *Strukturelle Rationalität,* Stuttgart: Redam 2001 and in *Decision Theory and Ethics,* München: Utz 2005, therefore I shall not repeat these arguments here.
17. cf. Leonard J, Savage, *The Foundations of Statistics*, New York 1954 and Richard C. Jeffrey, *The Logic of Decision,* New York 1965.
18. "Certain" in the sense that the subjective probability is I.
19. as it is in Savage's *Foundations of Statistics* , ibid.
20. cf. G.E.M. Anscombe, *Intention*, Oxford 1957.
21. It is unnecessary to demonstrate this mathematical deduction here, because our requirements of minimal coherence regarding deontologically motivated preferences are equivalent with the axioms of the utility theorem, see John von Neumann/Oskar Morgenstern, *Theory of Games and Economic Behavior*, Princeton 1947 and R. Duncan Luce/ HowardRaiffa, *Games and Decisions,* New York/London/Sydney 1957, chap 2
22. *Optimizes subjective value = maximizes expected subjective value.*

23. Since Frank P. Ramsey was the first who developed the idea, almost unnoticed, of the utility-theorem, I call the property of a preference-relation, which consists in meeting all the decision-theoreric axioms "Ramsey-Coherence". This terminology has the additional advantage that the emphasis is laid on coherence, which is what the axioms are about.
24. *Weighing Goods. Equality, Uncertainty, and Time,* Oxford 1991, chap. 5.
25. I am grateful for having had the opportunity to discuss my arguments with John Broome, Ned McClennen, Michael Slote, Wolfgang Spohn and others during the conference. I thank Martin Rechenauer and Thomas Schmidt for reading this paper and proposing valuable amendments.
26. See Nida-Rümelin (1997), and, more detailed, Nida-Rümelin (1993, Parts III and IV).
27. In fact, I take this dichotomy to be one of the inventions of modern ethical theory which should be given up (cf. Nida-Rümelin, 1998). For the argument of this paper, it is not necessary to go further into this matter.
28. Cf. Austin (1955).
29. Note that we do not have to subscribe to some kind of institutional analysis like the one presented in Searle (1969). It is not necessary to assume that the fact that X has uttered 1 which under some conditions means that X has promised to φ alone is sufficient for X having a good reason to φ. The discussion of the 'naturalistic fallacy' is not relevant for the point under discussion here.
30. One might suppose that with this description we opted implicitly for an outcome-structure of the *Assurance Game.* This is not the case, as I tried to show in Nida-Rümelin (1991).
31. There is a critical point here: Is this statement also true for the ideal theory of practical rationality? Or should we require a complete theory of good reasons for action to be such that weak incoherence of action-guiding preferences is incompatible with being adequately based on conative attitudes and being determined by adequate conative attitudes?
32. Those who argue that there are genuine moral dilemmas think that our moral reasons for action cannot be made coherent in this sense (cf. Gowans, 1987),
33. Contrary to the first three axioms the last four mentioned axioms make use of probability considerations and this might undermine their initial plausibility. I think they are equally plausible, but I cannot go into details here. I maintain that even a deontological framework can account for probability considerations (cf. Nida-Rümelin, 1993, § 51).
34. Cf. also McClennen (1990).
35. Sen (1997).
36. In fact I think that there is not only an underdetermination of theories by empirical data, there is also an underdetermination of application. This underdetermination can take the form of different possible interpretations of the basic terms. Those who argue against the compatibility assumption adhere to a more conventional interpretation as it is mostly used in economic and other social science applications of rational choice theory. In the end, it is theoretical fruitfulness of conceptual frameworks which decides.
37. Cf. Sen (1971, Chapters 1 and 1*), see also Sen (1977) and Kem/Nida-Rümelin (1994, Sect. 1.2–1.4).
38. Cf. Vallentyne (1988).
39. Cf. Broome (1991, Chapter 5.4).
40. I used that term already in my habilitation-thesis which was handed in December 1988 (under the title 'Der Konsequentialismus. Rekonstruktion und Kritik', LMU Munich, see esp. Chapter 4, pp. 417–479). A shortened version of this thesis was completed shortly after but published only four years later in 1993 because the book-series 'Ethica' which de Gruyter planned and in which the manuscript was scheduled to appear was cancelled. This is the reason why in Nida-Rümelin (1993) there is no reference to publications which appeared later than 1988 – e.g., Edward McClennen's work on *resolute choice* (cf. McClennen, 1990) or John Broome's idea of *weighing goods* (cf. Broome, 1991).
41. I am convinced that act-utilitarianism is not right in this, see Nida-Rümelin (1993, §§ 26–29 and 36).

42. I introduced the term "structural rationality" in my habilitation thesis (Nida-Rümelin (1993)) in order to describe phenomena of rational agency that are not covered by the consequentialist optimization model. Later on, I extended this account to a general nonconsequentialist theory of practical reason. Cf. Nida-Rümelin (2001), Nida-Rümelin (2005a) and Nida-Rümelin (2011). I mention these publications in order to make the context explicit, but the argument developed in this paper is self-contained. For a short presentation of the account of structural rationality, cf. Nida-Rümelin (1997b), followed by Nida-Rümelin (2000) and Nida-Rümelin (2005b). In the following, I shall refer exclusively to English publications.
43. Cf. Nida-Rümelin (1997a).
44. I borrow this example from Nagel (1970).
45. The Pareto condition requires that alternatives that are preferred by all people must also be preferred by the collective. Correspondingly, we find a Pareto efficient allocation of goods among a group of individuals if it makes at least one individual better off while making no one else in the group worse off.
46. Classical texts dealing with these universal traits of language are Lewis (1969; 1974, 23); Searle (1969, ch. 2.7); Davidson (1973). In the chapter "Communication" of Nida-Rümelin (2001), I argue that the constitutive rules of communication are not to be understood as solving coordination games, as Lewis assumes, but as solving cooperation games.
47. Regarding this point, my views differ from the position held by Tuomeia in Tuomela (2007, ch. 7).
48. Cf. Savage (1954).
49. Cf. Tuomela (2007, ch. 7), but much earlier, also Rescher (1975).
50. Cf. Margolis (1982) and Scheffler (1982).
51. Cf. Shapiro (2006).
52. Cf. Nida-Rümelin (1994).
53. Cf. Spohn (2012).
54. Preference relations are defined here in the usual way, i.e., they are reflexive, transitive, and connective.
55. This terminology, however, is misleading, because one of the most interesting applications of collective choice theory is the analysis of political decision making in which political preferences are aggregated. But the latter do not usually represent individual utilities. Classical works in social choice theory are Arrow (1963); Sen (1986).
56. Cf. Varian (1974); Pazner and Schmeidler (1978); Fleurbaey and Trannoy (2003).
57. Schelling (1960).
58. Kantian ethics can be read as incorporating the imperative to respect individual autonomy and refrain from individual optimization if there is a cooperative solution. Two possible readings of the categorical imperative underlie this interpretation. First, the generalization of maxims should be logically possible; second, individuals should be able to desire generalizing their maxims as a principle of general behavior. The second interpretation excludes maxims that are destructive for cooperation because no one desires that all citizens of a society act in a way that destroys cooperative behavior. However, the problem with Kantian ethics is that it has no answer to my (5). Hegelian practical philosophy contains a culturalist and historicist solution of (5), but this results in the loss of normativity and universality. Counterfactual rationality tries to keep as much normativity and universality as possible without losing the grounds of our *Lebenswelt.*
59. Amartya Sen has developed several arguments that speak in favor of a new decision-theoretic account that "drives a wedge between choice and consequences," but he never said what type of new decision theory, including game theory and collective choice, should result. The account of structural rationality can be read as an answer to Sens Challenge. Cf. Sen (1977, 1997) and Nida-Rümelin (2000).
60. Note that this case is not the same as the I-mode type of group action that Tuomela (2007).
61. This stands in contrast not only to collectivist and institutionalist notions like that of Mary Douglas (cf. Douglas (1986) and other publications of hers) but also to moderate accounts like that of Margaret Gilbert (cf. Gilbert (2000)).

Suffix

Rebecca Gutwald and Niina Zuber

The Meaning(s) of Structural Rationality – A Suffix[1]

In philosophy, questions about rationality and its normative character can be situated within two theoretical debates that have largely evolved separately from each other. First, rationality is a central topic in philosophy of action. Prominent authors such as Thomas Scanlon, Christine Korsgaard, Michael Bratman or Joseph Raz have debated why we should be rational, how rationality constrains our beliefs, intentions and attitudes, or, whether we have any reason to be rational at all. Second, rationality is a central concept in the discussion of so-called rational choice theory (RCT) in economics and classic game theory. A central question, which most authors in these areas answer in the affirmative, is whether rationality is purely instrumental or whether there are forms of rationality that cannot be explained by this standard model, which purports that rationality is nothing more than a tool to find the best means to one's end(s).

This text serves two purposes. First, we present Julian Nida-Rümelin's work on the concept of "structural rationality", which expands the instrumental model without denying its importance. He criticizes theories of reason that neglect or ignore the importance of instrumentality in several ways. We argue that his model of rationality is philosophically superior to both the dominant traditions in contemporary theories of action and rational choice since his approach explains phenomena like rules-commitment strategies, cooperation, collective action using the formal method of rational choice (which is difficult for many approaches). It is the only theory that takes into account what advocates of rational choice theories *as well as* theorists in philosophy of action have to say, and which brings these two – hitherto separate – discussions together.

A revised version of this chapter was published in ProtoSociology – An International Journal and Interdisciplinary Project, http://www.protosociology.de/on-philosophy.html, 2018.

© Springer Nature Switzerland AG 2019

J. Nida-Rümelin, *Structural Rationality and Other Essays on Practical Reason*,
Theory and Decision Library A: 52,
https://doi.org/10.1007/978-3-319-95507-0

We support our hypothesis, by, second, comparing Nida-Rümelin's account of structural rationality to the most prominent approaches to (structural) rationality in Anglo-Saxon philosophy and rational choice theory. We will show that most theorists have a more limited and simpler view than Nida-Rümelin's, because they often only focus on short-term, instrumental and individualist forms of rationality. They thus disregard two important considerations. First, they fail to see that the critical arguments that are brought forward against rational choice theory also apply to their theories. Second, they are unable to explain various and diverse forms of action and decision-making: cooperation, collective action, actions and practices that have value on their own (friendship, enjoying arts, sports). We will point out that Michael Bratman's account of rationality and, to some extent, Joseph Raz's are the only approaches that are able to address some of these problems, even though they still adhere to the individualist perspective too much. These are the reasons why we believe that it is highly worthwhile to engage more with Nida-Rümelin's account, which is inspired by Ludwig Wittgenstein, Immanuel Kant and Aristotle.

In what follows, we give an overview of Nida-Rümelin's account by showing how it differs from the term "Structural Rationality" as it is used in Anglo-Saxon philosophy of action, and compare it to the concept of rational choice in economic theory. We highlight how Nida-Rümelin's theory of structural rationality transcends these limitations of these two approaches and also bridges the gap between them.

In Anglophone philosophy of action, the expression "structural rationality" is commonly ascribed to Thomas Scanlon (2007). As we point out below, however, Nida-Rümelin has been using the term already earlier in his works on rationality written in German (most notably in Nida-Rümelin 2001). It is also important to highlight that the Scanlonian use of the term diverges strikingly from Nida-Rümelin's, since the phenomenon it describes differs in scope and meaning. It should rather be called, as Scanlon often does, the phenomenon of "structural irrationality". Scanlon and other philosophical theorists of action – for instance Bratman or Broome – examine it as the problem that some people fail to have their attitudes combined coherently, and thus structured, in a rational way. Thus, irrationality is not only located within one or several particular attitudes, but also in their combination or rather arrangement *as such*. These authors thus claim that it is irrational to combine certain attitudes.

Let us take the following example to illustrate this (Langlois 2014): You intend to spend the weekend with your family while simultaneously intending to complete your new manuscript by Sunday. You also believe that both of these goals are achievable. Intuitively, a person truly intending both things *at the same time* makes a mistake of rationality, which cannot be attributed to the content of these intentions. Hence, the structure of the set of attitudes is irrational. This explanation of irrationality is to be understood formally. Conversely, authors that espouse this theory hold, there must be structural requirements of rationality that create a normative demand on how people are supposed to combine their attitudes. Several questions arise from this conceptualisation and the theory of rationality that lies beneath it. Most fundamentally, there is the issue how standards of rationality have normative power over us. Does the tenet that a certain combination of attitudes is irrational

have normative traction? If so, how does it have to be characterised? What are the requirements for structural rationality of this kind? Obviously, they are located in conditions of coherence and consistency (Langlois 2014). Hence, one of the most fundamental questions regarding the conceptualisation of structural (ir)rationality pertains to the rules that determine in which way attitudes of an agent should be combined and arranged.

For Scanlon and his followers, rationality is of instrumental character in the way David Hume has described it, i.e. it pertains to the relationship between the means and ends for an action. Take again the example we have just described: the protagonist has two goals (writing an essay, and, spending time with her family). Neither of them is irrational per se, since any of them can be achievable within the means of the author. Also they are not in principle incompatible or contradictory in relation to each other. Yet, there is something problematic in holding them both at the same time, because they seem in tension to each other. It is, at the very least, very difficult to reach both of these goals simultaneously in the manner planned, if we confine the idea of rationality to an instrumental, subjectivist one as Scanlon et al. do. We cannot criticise the goals or the reasons behind them as irrational, since rationality is a matter of picking suitable means for one particular goal, e.g. finding time and space to write the paper if you want to write a paper or finding the right means of transport of reaching one's family. Viewing attitudes in this isolated and individualist way thus gives us no ground to criticise either the goals behind them or the way that they relate to each other, because this would transcend the standard instrumental, individualist model of rationality. What remains is to look at the ways that agents combine their attitudes, intentions and reasons. This, however, begs several questions, e.g. what kind of requirements regulate the combinations or where the normative traction of these requirements comes from.

Nida-Rümelin avoids these difficulties, since he neither assumes that rationality is purely and only instrumental nor that intentions and reasons or goals are merely subjective. For Nida-Rümelin, the structures that people adopt for themselves can be objective, and the sources of reason are plural, and also dependent on the *Lebensform* that people are embedded in. In our orientation to adopt and follow structures, we agents do more than simply looking for the best means based on reasons and intentions that relate to mere subjective and current goals.

To illustrate this, consider the following example: Imagine a father watching his son falling and hurting himself. The father hurries to his side and immediately consoles him. What is his reason for acting in this case? What is his goal? The father acts because he is father. He is doing this because consoling is part of his father-son-relationship: it is an essential part of such a relationship to help and be empathic. He does not need to have a specific, subjective reason or goal, like, say, that his son thanks him or that he will stop crying. This may be part of his motivation, but it does not tell the whole story. Actually, a father does not have to find a specific reason at all in this situation. Being a father can be a sufficient reason for action. However, this way of reason generating cannot be fully explained by the individualist and instrumental model of rationality.

Nida-Rümelin also diverges from the model of rationality explained above in several other ways. He accepts Scanlon's and Korsgaard's understanding of practical rationality as the capability of following a rule that allows us to arrange practical reasons coherently, since Nida-Rümelin emphasizes that any action – regardless of its reasons – must be Ramsey-coherent. Nida-Rümelin, however, rejects that the rule, according to which attitudes, desires, obligations etc. are aligned to, can be formulated in one way only: either as a universalizing maxime (e.g. Christine Korsgaard), consequentialist (e.g. homo oecnomicus-models, Garry Becker), virtues (e.g. Phillipa Foot, Rosalind Hursthouse), social role principles (*homo sociologicus*-models). To arrange the plurality of our practical reasons we need more than just one standard, and this is exactly the point where Nida-Rümelin's approach comes into play.

Furthermore, this leads into the question where reasons are located. For example, Korsgaard highlights that reasons arise out of our practical identity whereas Nida-Rümelin adopts a realist perspective: Reasons do not arise out of practical identities we choose but out of the form of life we inevitably find ourselves in: in father-son-relationships, friendships, professional roles, virtues etc. A theory of practical rationality needs to take pluralism seriously and not reduce rationality to either only one accepted motive (e.g. self-interest) or to only one rule (e.g. universalism).

Comparing the Scanlonian account of strucutural rationality to Nida-Rümelin's, we can see that they do not only form a different conceptualisation of rationality itself, but also that the characterisation of structures deviates strikingly. Most importantly, Nida-Rümelin's account is objectivist, but not (reason)-foundational. Structures are not (only) ways that an individual's attitudes have to each other, but can rather be located in the world that surrounds an individual. An agent can decide to adopt structures for herself, but she can neither create them solely on her own nor can she modify them significantly without violating it. Hence, a person who does not act like a father in the example above, even though he claims to be one (e.g. somebody that hides from the kid) also commits a violation of rationality in the structural sense. Hence, Nida-Rümelin's account and the Scanlonian approach to structural rationality only share a limited number of features. Despite bearing the same label, they should not be confused, and we hope to have illustrated why Nida-Rümelin's account is richer, more pluralistic and more comprehensive than Scanlon's, Korsgaard's and similar approaches.

Apart from these approaches, we find some other accounts in theory of action that are more similar to Nida-Rümelin's, mostly because they are critical of the model of instrumental rationality (although it must be noted that none of them is equally comprehensive). Joseph Raz, for instance, evaluates what he calls "the myth of instrumental rationality" (2005) critically, and claims that instrumental reasons cannot fully be accounted for by the prominent theories of instrumental rationality. Joseph Raz drives a wedge between the ideas of an agent's goals or rather intentions and her reasons. In other words: intentions and goals do not automatically generate reasons for action. He thus questions the foundational idea of the normativity of reasons. He also introduces a wider net of considerations that we have to take into account when we aim to understand actions, such as the processes that lead to an

intention and action, and with it the underlying habits, dispositions and patterns of thinking. In this way, Joseph Raz also criticizes the most prevalent views on instrumental rationality as insufficient and introduces similar considerations as Nida-Rümelin's.

Joseph Raz also states that Michael Bratman's account offers a prominent and comprehensive account of rationality that avoids the problems of many accounts. Michael Bratman's "planning theory of action" comes close to Nida-Rümelin's account, because it shares two central features with Nida-Rümelin's theory. First, Bratman acknowledges that there are social rationality norms for intentions and policies (Bratman 2014): friendship, love, dancing together, conversations etc. are all practices and forms of action that cannot be explained without references to sociality. Second, Bratman refers to "structures" (ibid. p. 8) albeit in a different sense than Nida-Rümelin: intentions should be seen as plan states that occupy space in coordinating plans that structure agency in a diachronic way. Thus, intentions have a coordinating role in providing continuity and organization over time and order conduct. Intentions in Bratman's sense thus share some characteristics with Nida-Rümelin's account, but Bratman's account has one significant shortcoming: Bratman does not say much about the sources of reason and the question whether structures are objective. His main interest lies in the specification of plans as part of individual and social action, but he does not proceed to explain the bigger picture, namely what the nature of structures is in a world of agents.

Another unique feature of Nida-Rümelin's account is that he refers to another prominent account of rationality that theorists of actions are usually silent about: rational choice theory in economics. Rational choice theory is the best example of ignoring facets of human practical reasoning and thus drawing a picture of human actions that does not correspond with our daily experiences, especially pointed in the theory of *homo oeconomicus*. It is precisely the idea of an agent only maximizing her preferences that simply fails to acknowledge the plurality of human life. The shortcomings of rational choice theory are linked to its understanding of rationality solely as instrumental: an action is considered as rational only if it fits best to achieve the aim, i.e. satisfies the desire best. Analyzing and explaining actions in terms of instrumental reason only, however, is clutching at straws. Either rationality is taken as instrumental or as prudent. Other forms of actions are neglected, or rather are deemed irrational. Reducing all practical reasons to maximizing strategies just leads into an artificial language that does not picture reality.

The empirical results of behavioral economics reveal the problems of such a narrow concept of rationality: human behavior is not always in accordance with the theory of practical rationality understood as instrumental or acting prudently as Sunstein, Kahneman and Tversky show. However, the critique of behavioral economists on RCT remains superficial since it moves along with the assumptions of instrumental rationality. They try to nudge human behavior to fit practical rationality, but they do not scrutinize their theory. This is not only true for behavioral economics, but also for psychology and the concept of bounded rationality (e.g. Simon, Gigerenzer).

Practical reasoning needs to encompass the totally of our form of life with its specific justificatios of actions. A theory that fails in explaining various aspects of our daily actions distorts our daily praxis of practical reasoning, leads to a denial of reality, and thus leads to detached theories of human actions. A comprehensive theory of structural rationality must explain the dialectic between a self-inventing agent and given structures of form of lives. However, both the agent as well as the form of life should not be mistakenly assumed to be independently given. A theory of action that does not entail all kinds of aspects of human agency cannot give an appropriate account of practical rationality.

Nida-Rümelin's approach is thus comprehensive and radical: It transcends the reductive narrowness of instrumental rationality without denying its practical impact. Actions do exits that are outlined in accordance to utility maximizing or even self-interest maximizing. Yet, not all actions are to be understood in these terms. Actions that are oriented toward social roles, for example, cannot count as irrational only because no underlying maximizing heuristics is found. The concept of bounded rationality tries to embed instrumental rationality into a form of life to highlight limits of our cognitive capabilities and selective perceptions. However, the agent is still left within the realm of cost-benefit-reasoning. The idea of social preferences (e.g. Charness and Rabin) or meta-preferences (Sen) cannot encompass the plurality of human actions. According to Nida-Rümelin they ignore the plurality of reasons that drive agency. Hence, they coerce agency in fitting into a theory that undermines humanity. His theory of structural rationality, however, acknowledges daily patterns of interaction and meaning.

To illustrate this, consider once more the example of the father described above: the father does not act because he wants to maximize his utility. He does not need to analyze costs vs. benefits to find a reason for acting like a father. Being a father can be a sufficient reason for action, and we do not need any kind of cost-benefit-analysis. There might be agents that act also as fathers according to a cost-benefit-reasoning, but most of them would deny that.

For Nida-Rümelin it is of great importance that some space is left for practical deliberation in practical life: An agent can evaluate her desires, interests, obligations, roles etc. and become author of her life. She will determine herself, what reasons will be decisive for her actions that make up her life, i.e. that will structure her life. Thus, there exists a deliberation equilibrium of individual and structural reasons of actions, which finally enable –to speak in Nida-Rümelin's terminology –a coherentist lifestyle. This lifestyle is considered as rational: an agent that lives within a plurality of reasons that outline her life in such manner that she is able to create a compact, upright, and authentic life. Nida-Rümelin's structural rationality enables to think of an agent with integrity within a plurality of reasons and forms of lives.

References

Bratman, M. 2014. *Shared Agency: A Planning Theory of Acting Together.* Oxford: Oxford University Press

Charness, C., and M. Rabin. 2002. Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics* 117(3): 817–869.

Gigerenzer, G., and R. Selten (Hg.). 2002. *Bounded Rationality. The Adaptive Toolbox. Dahlem Workshop on Bounded Rationality*, 1st MIT Press pbk. ed. Cambridge, MA: MIT Press.

Korsgaard, C. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press

Jolls, C., C.R. Sunstein, and R.H. Thaler. 2000. A Behavioral Approach to Law and Economics. In *Behavioral Law and Economics*, Cambridge Series on Judgement and Decision Making, ed. Cass R. Sunstein, 13–59. Cambridge University Press.

Langlois, D.J. 2014. *The Normativity of Structural Rationality.* Doctoral dissertation, Harvard University, online http://nrs.harvard.edu/urn-3:HUL.InstRepos:13067678. Accessed on 24 Apr 2018.

Nida-Rümelin, J. 2001. *Strukturelle Rationalität – ein philosophischer Essay über praktische Vernunft.* Ditzingen: Reclam

Raz, J. 2005. The Myth of Structural Rationality. *Journal of Ethics and Social Philosophy* 1(1): 1–20.

Scanlon, T. 2007. Structural Irrationality. In *Common Minds: Themes From the Philosophy of Philip Pettit*, G. Brennan, R. Goodin, F. Jackson, and M. Smith, 84–103. Oxford: Clarendon Press.

Simon, Herbert A. 2008 (posthumously). *Economics, Bounded Rationality and the Cognitive Revolution*. Edward Elgar Publishing

Sen, A. 2003. *Rationality and freedom*, 2. printing. Cambridge, MA: Belknap Press of Harvard University Press.

Sen, A. 1977, Summer. Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy & Public Affairs* 6(4): 317–344

Tversky, A., and D. Kahneman. 1983. Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review* 90(4): 293–315. *reprinted in* Gilovich, Thomas, Dale Griffin, and Daniel Kahneman (ed.). 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment*, 19–48. Cambridge: Cambridge University Press.

Note

1. A revised version of this article has been published by R. Gutwald and N. Zuber in *ProtoSociolgy – An International Journey and Interdisciplinary Project,* Rubrum – On Contemporary Philosophy. 2018. http://www.protosociology.de/on-philosophy.html, accessed on 8.10.2018.